OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

THE OXFORD HANDBOOK OF

APPROACHES TO LANGUAGE EVOLUTION

9780192886491_Book.indb 1 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:13

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

OXFORD HANDBOOKS IN LINGUISTICS

RECENTLY PUBLISHED

THE OXFORD HANDBOOK OF LANGUAGE AND RACE
Edited by H. Samy Alim, Angela Reyes, and Paul V. Kroskrity

THE OXFORD HANDBOOK OF LANGUAGE PROSODY
Edited by Carlos Gussenhoven and Aoju Chen

THE OXFORD HANDBOOK OF LANGUAGES OF THE CAUCASUS
Edited by Maria Polinsky

THE OXFORD HANDBOOK OF GRAMMATICAL NUMBER
Edited by Patricia Cabredo Hofherr and Jenny Doetjes

THE OXFORD HANDBOOK OF COMPUTATIONAL LINGUISTICS
Second Edition
Edited by Ruslan Mitkov

THE OXFORD HANDBOOK OF THE MENTAL LEXICON
Edited by Anna Papafragou, John C. Trueswell, and Lila R. Gleitman

THE OXFORD HANDBOOK OF ETHIOPIAN LANGUAGES
Edited by Ronny Meyer, Bedilu Wakjira, and Zelealem Leyew

THE OXFORD HANDBOOK OF EXPERIMENTAL SYNTAX
Edited by Jon Sprouse

THE OXFORD HANDBOOK OF WORD CLASSES
Edited by Eva van Lier

THE OXFORD HANDBOOK OF IRISH ENGLISH
Edited by Raymond Hickey

THE OXFORD HANDBOOK OF SOUTHEAST ASIAN ENGLISHES
Edited by Andrew J. Moody

THE OXFORD HANDBOOK OF THE FRENCH LANGUAGE
Edited by Wendy Ayres-Bennett and Mairi McLaughlin

THE OXFORD HANDBOOK OF VOWEL HARMONY
Edited by Nancy A. Ritter and Harry van der Hulst

THE OXFORD HANDBOOK OF RITUAL LANGUAGE
Edited by David Tavárez

THE OXFORD HANDBOOK OF ARCHAEOLOGY AND LANGUAGE
Edited by Martine Robbeets and Mark Hudson

THE OXFORD HANDBOOK OF LANGUAGE AND PREJUDICE
Edited by Jane Setter, Sender Dovchin, and Vijay A. Ramjattan

THE OXFORD HANDBOOK OF APPROACHES TO LANGUAGE EVOLUTION
Edited by Limor Raviv and Cedric Boeckx

For a complete list of Oxford Handbooks in Linguistics please, see pp. 607–638

9780192886491_Book.indb 2 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:13

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

THE OXFORD HANDBOOK OF

APPROACHES TO LANGUAGE EVOLUTION

Edited by

LIMOR RAVIV

and

CEDRIC BOECKX

OXFORD
UNIVERSITY PRESS

9780192886491_Book.indb 3 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:14

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

© editorial matter and organization Limor Raviv and Cedric Boeckx 2025
© the chapters their several contributors 2025

The moral rights of the authors have been asserted

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, transmitted, used for text and data mining, or used for training artificial intelligence, in any form or by any means, without the prior permission in writing of Oxford University Press, or as expressly permitted by law, by license or under terms agreed with the appropriate reprographics rights organization. Inquiries concerning reproduction outside the scope of the above should be sent to the Rights Department, Oxford University Press, at the address above.

You must not circulate this work in any other form
and you must impose this same condition on any acquirer

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data

Data available

Library of Congress Control Number: 2024947192

ISBN 978–0–19–288649–1

DOI: 10.1093/oxfordhb/9780192886491.001.0001

Printed and bound by
CPI Group (UK) Ltd, Croydon, CR0 4YY

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Contents

List of figures ix
List of Contributors xvii
Introduction xxiii
Limor Raviv and Cedric Boeckx

PART I. HUMAN DATA

1. Iterated learning experiments 3
Monica Tamariz and Aliki Papa

2. Artificial language learning 25
Shira Tal, Inbal Arnon, and Jennifer Culbertson

3. Communication experiments: Social interaction in the formation of novel communication systems 41
Thomas F. Müller and Limor Raviv

4. Virtual reality as a tool to study language evolution 63
Jonas Nölle and David Peeters

5. Studying child-directed speech 83
Vera Kempe and Marisa Casillas

6. Corpus-based approaches to evolutionary dynamics in language 111
Stefan Hartmann and Michael Pleyer

7. Adaptation 127
Caleb Everett

8. The evolution of the grammar of the body in sign languages 143
Wendy Sandler, Svetlana Dachkovsky, and Rose Stamp

9. Silent gesture: Gesture studies with hearing participants 163
Danielle Naegeli and Marieke Schouwstra

10. Methods to study evolution of iconicity in sign languages 177
Anita Slonimska and Asli Özyürek

9780192886491_Book.indb 5 /12_first_proofs/first_pr[illegible]/xml_for_typesetting 27-Mar-25 23:02:14

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

11. The study of Creole languages 195
Damian Blasi, Felicia Bisnath, and Pui Yiu Szeto

12. Palaeogenomics: A window into the genetic basis of derived traits in *Homo sapiens* 209
Juan Moriano, Cedric Boeckx, and Martin Kuhlwilm

13. Joint analysis of human linguistic and genomic variation 227
Yakov Pichkar and Nicole Creanza

14. Studying long-term evolutionary processes over short timescales: Methods in cumulative cultural evolution 241
Maxime Derex and Charlotte Brand

15. An integrative approach to early symbolic evolution: Experimenting with the past 259
Isobel Wisher and Kristian Tylén†

PART II. SIMULATIONS

16. The Bayesian iterated learning model 277
Vanessa Ferdinand

17. Communication games: Modelling language evolution through dyadic agent interactions 301
Xenia Ohmer and Christine Cuskley

18. Computational methods for language variation and convergence 319
Katie Mudd and Bart de Boer

19. Language evolution with deep learning 335
Mathieu Rita, Paul Michel, Rahma Chaabouni, Olivier Pietquin, Emmanuel Dupoux, and Florian Strub

20. Task-driven language evolution with swarm robotics 371
Nicolas Cambier and Roman Miletitch

21. Linguistic evolution in time and space: Addressing the methodological challenges 391
Chundra Cathcart and Balthasar Bickel

22. Information-theoretic and machine learning methods for semantic categorization 423
Francis Mollica and Noga Zaslavsky

9780192886491_Book.indb 6 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:14

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

23. Causal graphs as a tool for exploring language evolution 441
Kateryna Krykoniuk and Seán G. Roberts

PART III. ANIMAL DATA

24. Observational work 457
Marco Gamba and Chiara De Gregorio

25. Re-imagining great ape gesture (coding) 469
Erica A. Cartmill, Emilie Genty, Kirsty E. Graham, Charlotte Grund, and Catherine Hobaiter

26. Evolutionary roots of human cooperative communication: Using the CA-assisted comparative approach for quantitative and qualitative analyses 487
Simone Pika and Ray Wilkinson

27. Multi-level approaches to assessing flexibility in great ape communication 503
Marlen Fröhlich and Carel P. van Schaik

28. Methods of examining the processing and production of rule-based sequences in animals 523
Mélissa Berthet, Maël Leroux, Simon W. Townsend, and Stuart K. Watson

29. Bioacoustics and rhythm 539
Daria Valente and Andrea Ravignani

30. Hominin tool evolution and its (surprising) relation to language origins 553
Ronald J. Planer, Elisa Bandini, and Claudio Tennie

31. Comparative neuroimaging to study the neural infrastructure for language 577
Nicole Eichert and Rogier B. Mars

32. How to study animal syntax 593
Toshitaka N. Suzuki

Index 613

9780192886491_Book.indb 7 /12_first_proofs/first_pr…/xml_for_typesetting 27-Mar-25 23:02:14

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

9780192886491_Book.indb 8 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:14

FIGURES

1.1 In this example iterated learning chain, the outputs change in response to cognitive biases. 4

1.2 Example meaning spaces from iterated learning experiments. 6

1.3 Two seed languages. 7

1.4 Three output languages produced at the last generation of iterated learning chains. 7

1.5 Two languages that exhibit systematicity but not compositionality. 8

1.6 Population structure in iterated learning chains. 9

1.7 Screen captures from the procedure in Kirby et al.'s (2015) study, which included a collaborative, communicative testing task. 11

1.8 Some choices in transmission-chain iterated learning experimental design. 12

1.9 Mantel test of systematicity in a subset of signals and meanings in the output language from Kirby et al. (2008) shown in Figure 1.4b. 13

1.10 Calculation of *RegMap* of the last-generation language from Kirby et al. (2015) shown in Figure 1.4b. 14

1.11 A Pareto frontier contains the area of possible languages. Languages that are on the frontier line optimize two desirable properties: simplicity and expressivity. 17

1.12 Regularity of the mappings (partial *RegMap*) between the two signal segments and two meaning dimensions over generations in a chain (based on Kirby et al. 2015). 18

2.1 Stimuli used in Culbertson et al. (2020) to denote novel objects that participants are unfamiliar with. 31

3.1 A basic referential communication task. 42

3.2 Case study on shared visual context. 50

3.3 Case study on the role of social structure using a group communication paradigm. 51

4.1 (a) Example of a user wearing a VR headset to immerse himself into a virtual world. (b) Example of an experimental participant being immersed in a virtual forest in a CAVE, while her electrophysiological brain activity is continuously recorded via an electroencephalogram (EEG). 64

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

4.2 (a) Traditionally, ecological validity and experimental control have been conceptualized as opposite ends of a continuum. (b) In contrast, different experimental paradigms are represented as points in a three-dimensional coordinate system, where ecological validity (x-axis), experimental control (y-axis), and immersion (z-axis) form orthogonal aspects of experimental design. 66

4.3 Chimpanzee touchscreen setups. 72

4.4 (a) The two virtual environments in the study by Nölle, Kirby et al. (2020). (b) Proportions of FoR strategy in both environmental conditions (Experiment 1) and (c) over time across both experimental blocks (Experiment 2). (d) The novel communication medium used in the experimental semiotics version of OrbHunt: the Director could select one of eight shapes on their controller and send them to appear in the Seeker's field of view. (e) Performance of dyads grounding a novel communication system. 74

7.1 Two world maps, each with a raster layer based on bioclimatic variable six from the WorldClim dataset (Fick and Hijmans 2017). 134

7.2 Average jitter rates (see Formula 1) of recordings in 997 locations in China, in accordance with deciles of average ambient water vapour pressure of those locations. 137

7.3 Labiodental usage across 2,729 dialects, categorized according to dominant subsistence strategy of the associated culture. 138

8.1 The Grammar of the Body: some of the linguistic properties conveyed by actions of particular articulators of the body. 145

8.2 Iconic signs in a typical SOV sentence, MAN DISH WASH. 148

8.3 The boundaries of two clauses with conditional meaning: SCORPION-STING/NO-EFFECT—'If a scorpion stings the child, there will be no effect'. 149

8.4 Prosody marks a conditional sentence in a younger second-generation signer of ABSL. REFUSE/NOTHING-TO-DO. 'If he refuses, then there's nothing I can do'. 149

8.5 Recruiting additional articulators adds complexity as a sign language emerges. Data from ABSL. 150

8.6 A complex sentence with a relative clause in ISL, '[The little dog [that I found last week]] ran away'. 151

8.7 Three prosodic chunks in the ISL older signer's response corresponding to the nominal 'the man', the specifying predication 'who is driving a car', and the main predication 'is smoking' in the sentence 'The man who is driving a car is smoking'. The intonational signals are co-temporal only with the topic of the utterance, 'man'. 153

9780192886491_Book.indb 10 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:15

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

8.8 The same string in a younger ISL signer. The nominal and the relative clause, 'The man who is driving a car' constitute one prosodic constituent, marked by the intonational signals, forward head movement and squint. The main predication, 'is smoking', is separated prosodically from the preceding material by a rhythmic break (marked by a dotted line), as well as by an across-the-board change in intonational signals of the face and head. 154

8.9 Changes in the alignment of forward head movement (FHM) and squint with relative clauses across the three age groups of ISL signers. 155

8.10 Example of RST structure produced by an older ISL signer, analysed conceptually, with higher levels shown at the top and lower levels shown at the bottom. 156

8.11 Older signer: a higher-level boundary (Level 1, Summary) is signalled through the use of a lexical marker THAT'S-IT as well as head and torso movements. At the lower-level boundary (Level 6, Sequence), the signer also marks the breaks between three related sequenced expressions with head and torso movements. 157

8.12 Younger signer: The higher-level boundary (Level 2, Circumstance) (after EVERYONE) is signalled by multiple cues including a change of torso position, head thrust, and blink. At the lower level (Level 5, Cause), the signer explains: 'I got the paper and I really wanted (to sign up for) the veterinary' // 'because I always wanted to'. The boundary (after VETERINARY) is marked by smaller and less prominent cues: head movement and manual duration (hand hold). 158

9.1 Examples of stimulus pictures for different event types. 166

10.1 Different types of iconic mapping used for SPOON. 180

10.2 Lexical signs for TEA in a) Turkish Sign Language, b) Japanese Sign Language, and c) Italian Sign Language. 182

10.3 An iconic and simultaneous construction in Italian Sign Language of a cat gazing at, holding, and petting a bear. 186

10.4 Example of five *information density levels* for the referent pair: 'bird' and 'bunny'. 187

10.5 Descriptions of an image (*information density level 4*) produced by a) a hearing participant using silent gesture and b) a signer using Italian Sign Language. 188

11.1 Developmental cline. 202

13.1 The steps of a Procrustes analysis to compare two sets of data. 229

13.2 Schematic of the workflow from processing raw data to applying a Procrustes analysis. 232

9780192886491_Book.indb 11 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:15

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

13.3 Illustration of residuals. 234

13.4 A Procrustes rotation of the first two principal components of phonemic variation among Niger-Congo languages. 234

14.1 Diagram of Morgan et al.'s flint-knapping transmission-chain design. 247

14.2 Experimental task of Derex and Boyd (2015) totem pole experiment. 249

14.3 Game interface. 250

16.1 Schematic overview of Bayesian iterated learning. 280

16.2 The main components of Bayesian iterated learning in matrix representation. 282

16.3 Example code in R that implements Bayesian iterated learning. 282

16.4 Example code in R showing basic numerical and analytical solutions to the Bayesian iterated learning model. 284

18.1 The steps of the language game with an example. 323

18.2 Visualization of the steps and parameters in the agent-based model. 325

18.3 The mean lexical variability over the 2,000 model stages for a hundred repetitions of a run with ten agents being assigned to different groups depending on the run. 326

18.4 Example of an association matrix between forms and meanings. 328

19.1 Iterative learning process in a machine learning problem. 337

19.2 Scheme of a two-player communication game. 338

19.3 Examples of Lewis and negotiation games. 339

19.4 Attempts to go beyond Lewis and negotiation games by embodying agents into a two-dimensional world. 340

19.5 General view of a communicative agent. 341

19.6 Scheme of a neural network and the operations between two layers. 342

19.7 Classic neural network modules. 343

19.8 Summary of neural communicative agents modelling in a unidirectional communication game. 346

19.9 One-dimensional representation of the effect of adjusting the learning rate. 348

19.10 Training regimes for a one-dimensional regression problem trained with the Mean-Squared Error Loss. 350

19.11 Updates cycle of a neural network in supervised learning. 351

19.12 Scheme of the training loop of a communication game optimized with reinforcement learning. 354

19.13 In the Visual Discrimination Game, the Sender sees an image (sender observation) and communicates about it to the receiver. 356

9780192886491_Book.indb 12 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:15

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

19.14 Examples of neural communicative agents in the Visual Discrimination Game. 357

20.1 Swarm robotic experimental platforms. 373

20.2 Example of two Kilobot robots playing the MNG. 377

20.3 The evolution over time of the number of interactions between robots, the number of MNGs conducted, and the value of in- and out-group prosociality. 380

20.4 Probabilistic Finite State Machine guiding the foraging behaviour of our robots. 382

20.5 Evolution of the rate of collected resource items over time (average per nest) for different item distributions, item availability, and initial prosocialities. The nests have the same initial prosocialities. 383

20.6 Evolution of the rate of collected resource items over time (average per nest) for different item distributions, item availability, and initial prosocialities. 383

21.1 Illustration of the computation of rate likelihoods under the pruning algorithm (Felsenstein 1981) for a binary character and a hypothetical history of a binary character, visualized as a stochastic character map. 396

21.2 Binary continuous-time Markov process representing changes between presence and absence of a feature. 403

21.3 Simulated CTM processes showing transitions between states of a binary feature under different speeds ($s \in \{5, 10\}$) and stationary probabilities ($p \in \{0.1, 0.9\}$). 403

21.4 Austronesian languages in PHOIBLE. 407

21.5 Posterior samples of phylogeographic trajectories inferred by our model. 410

21.6 95% and 85% posterior credible intervals (CIs) for β_0^s and β_0^p. 411

21.7 Median branch-level dispersals for a tree from the tree sample, visualized using ggtree (Yu et al. 2018). 412

21.8 95% and 85% posterior credible intervals for segment-level deflections to the global effect of dispersal on speed of change and stationary probability of presence. 413

22.1 Application of Shannon's communication problem to semantic categorization, as proposed by Zaslavsky et al. (2018). 425

22.2 a) Instantiation of the communication model in the domain of colour, where each colour is mentally represented by a probability distribution over a perceptual colour space; b) the chips in the Munsell colour chart; c) the communicative need *p(m)* is reflected in the shading of each colour chip; d) CIElab colour space resembling speaker intentions *m(u)* for each of the colour chips in the Munsell colour chart. 430

9780192886491_Book.indb 13 /12_first_proofs/first_pr[illegible]/xml_for_typesetting 27-Mar-25 23:02:15

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

22.3 Complexity–accuracy values for all languages in the World Color Survey (Cook et al. 2005), as well as English (Lindsey and Brown 2014), according to IB and Regier et al.'s (2015) formalization of the trade-off. 432

22.4 An example of a reference game played by a multi-agent system. 434

23.1 The basic elements of causal graphs: nodes and edges. 443

23.2 An example of a causal model. 444

23.3 Other conventions of causal graphs. 445

23.4 A causal graph of work by Raviv and colleagues linking population structure to aspects of language learning. 448

23.5 Dunbar's theory of language evolution in the context of other theories. 450

24.1 The BORIS interface. 461

24.2 Scan Sampling options in BORIS. 462

24.3 BORIS can export observations as Praat textgrids. 462

24.4 Time budget generated by BORIS. 464

24.5 Flow diagrams in BEHATRIX. 465

24.6 Levenshtein distance calculation in BEHATRIX. 465

25.1 Touching on contact gestures. 475

25.2 Example of ELAN coding of a bonobo interaction involving a multichannel combination (vocalization + gesture) to request food. 478

26.1 Turn 1 of the sequence: Mwea (L) REACHES OUT towards the meat. 494

26.2 Turn 2 of the sequence: Morten (R) turns away from Mwea in response to his request. 495

26.3 Mwea's (L) reaction to Morten's response: a bared-teeth display. 496

26.4 Turn 1 of the sequence: Mwea (L) REACHES towards the meat. 496

26.5 Turn 2 of the sequence: Morten (R) turns away and moves away the meat. 497

26.6 Turn 1 of the sequence: Mwea (R) requests meat with a PALM-UP gesture. 497

26.7 Turn 2 of the sequence: Morten (L) turns away and moves the meat away from Mwea. 498

26.8 Turn 1 of the sequence: Morten (R) OFFERS the meat to Mwea. 499

27.1 Cumulative number of identified signals in (a) Bornean and (b) Sumatran orangutans over consecutively observed interactions. 512

27.2 Overview of signals observed in wild and/or captive research settings. 512

28.1 Schematic representations of a) a first-order Markov chain, where the occurrence of element C only depends on the occurrence of the previous element B, and b) a second-order Markov chain, where the occurrence of the element C depends on the occurrence of the previous elements AB. 530

9780192886491_Book.indb 14 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:15

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

28.2 Schematic representations of a prediction suffix tree, where circles represent elements of a sequence, and arrows and their associated numbers represent the probability that an element follows another one. 531

29.1 Potential, non-exhaustive representation of rhythmic patterns within a single individual and their potential relationship to other individual rhythmic patterns in a chorus. 546

29.2 Methods (right side) used to test between possible features of a pattern (left side) and potentially place that rhythmic pattern in the taxonomy of Figure 29.1. 547

31.1 New developments in comparative neuroimaging. 581

31.2 Summary of human brain adaptations supporting speech and language. 587

32.1 Investigation of call production. 595

32.2 Japanese tits combine alert and recruitment calls into alert-recruitment call sequences when gathering other flock members to approach and harass stationary predators, such as shrikes. 596

32.3 Schematic representation for playback experiments. 597

32.4 Japanese tits move their heads horizontally in response to playback of alert calls, approach the speaker in response to recruitment calls, and combine these two behaviours in response to alert–recruitment call sequences. 598

32.5 Japanese tits discriminate between naturally ordered alert–recruitment call sequences and artificially reversed recruitment–alert call sequences. 599

32.6 Japanese tits mob a shrike specimen in response to alert–recruitment call sequences broadcast from a single speaker. 600

32.7 Japanese tits exhibit similar responses to playbacks of their own alert–recruitment call sequences (A-R) and artificially generated, mixed-species call sequences composed of conspecific alert calls and heterospecific recruitment calls (A-Rh). 601

9780192886491_Book.indb 15 /12_first_proofs/first_pro/xml_for_typesetting 27-Mar-25 23:02:15

Contributors

Inbal Arnon, Professor of Psychology, Department of Psychology, The Hebrew University of Jerusalem

Elisa Bandini, Postdoctoral Researcher, Department of Evolutionary Biology and Environmental Studies, University of Zurich

Mélissa Berthet, Postdoctoral Research Fellow, Department of Evolutionary Anthropology and Center for the Interdisciplinary Study of Language Evolution, University of Zurich

Balthasar Bickel, Professor, Department of Comparative Language Science and Center for the Interdisciplinary Study of Language Evolution, University of Zurich

Felicia Bisnath, Postdoctoral Researcher, Western Norway University of Applied Sciences

Damian Blasi, Research Professor, Catalan Institution for Research and Advanced Studies; and Professor, Center for Brain and Cognition, Pompeu Fabra University

Cedric Boeckx, ICREA Research Professor, Catalan Institute of Research and Advanced Studies and University of Barcelona

Charlotte Brand, Postdoctoral Research Fellow, Human Behaviour and Cultural Evolution Group, University of Exeter, and School of Psychology, University of Sheffield

Nicolas Cambier, Guest Researcher, Computational Intelligence Group, Vrije Universiteit Amsterdam

Erica A. Cartmill, Professor of Anthropology, Cognitive Science, and Animal Behavior, Indiana University

Marisa Casillas, Assistant Professor, Department of Comparative Human Development, University of Chicago

Chundra Cathcart, Senior Researcher, Department of Comparative Language Science and Center for the Interdisciplinary Study of Language Evolution, University of Zurich

Rahma Chaabouni, Research Scientist, Google DeepMind

Nicole Creanza, Assistant Professor of Biological Sciences, Department of Biological Sciences and the Evolutionary Studies Initiative, Vanderbilt University

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Jennifer Culbertson, Professor of Experimental Linguistics, Centre for Language Evolution, University of Edinburgh

Christine Cuskley, Reader in Language and Cognition, Language Evolution, Acquisition and Development Group, Centre for Behaviour and Evolution, Newcastle University

Svetlana Dachkovsky, Gordon Academic College, Haifa, and Sign Language Research Lab, University of Haifa

Bart de Boer, Professor, AI Lab, Vrije Universiteit Brussel

Chiara De Gregorio, Research Fellow, Department of Psychology, University of Warwick, and Postdoctoral Researcher, Department of Life Sciences and Systems Biology, University of Torino

Maxime Derex, Research Scientist, CNRS, Institute for Advanced Study in Toulouse, University of Toulouse Capitole

Emmanuel Dupoux, Directeur d'Etudes, EHESS, and Research Scientist, META

Nicole Eichert, Sir Henry Wellcome Postdoctoral Fellow, Wellcome Centre for Integrative Neuroimaging and Centre for Functional MRI of the Brain, Nuffield Department of Clinical Neurosciences, University of Oxford

Caleb Everett, Professor, Department of Anthropology and Department of Psychology, University of Miami

Vanessa Ferdinand, Senior Research Fellow in Computational Cognitive Science, Melbourne School of Psychological Sciences, University of Melbourne

Marlen Fröhlich, Freigeist Research Group Leader, Paleoanthropology, Institute for Archaeological Sciences, Department of Geosciences, University of Tübingen

Marco Gamba, Faculty Associate Professor, Department of Life Sciences and Systems Biology, University of Turin

Emilie Genty, Comparative Cognition Lab, University of Neuchâtel

Kirsty E. Graham, Wild Minds Lab, School of Psychology and Neuroscience, University of St Andrews, and Department of Psychology, Hunter College, CUNY

Charlotte Grund, Wild Minds Lab, School of Psychology and Neuroscience, University of St Andrews

Stefan Hartmann, Professor, Heinrich Heine University Düsseldorf

Catherine Hobaiter, Wild Minds Lab, School of Psychology and Neuroscience, University of St Andrews

Vera Kempe, Professor and Chair of Psychology of Language Learning, Abertay University

AQ: "Svetlana Dachkovsky, Gordon Academic College, Haifa, and Sign Language Research Lab, University of Haifa" --- This is different from the description given by the author during the affiliation management, which is: 'Research Associate, Department of Communication Sciences and Disorders, [University of Haifa]'. Which is correct, or should

9780192886491_Book.indb 18 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:15

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Kateryna Krykoniuk, Lecturer, School of English, Communication, and Philosophy, Cardiff University

Martin Kuhlwilm, Assistant Professor, Department of Evolutionary Anthropology and Department of Human Evolution of Archaeological Sciences, University of Vienna

Maël Leroux, Associate Professor, CNRS/EthoS, University of Rennes, and University of Caen Normandie

Rogier B. Mars, Professor of Neurosciences, Wellcome Centre for Integrative Neuroimaging and Centre for Functional MRI of the Brain, Nuffield Department of Clinical Neurosciences, University of Oxford, and Principal Investigator, Donders Institute for Brain, Cognition, and Behaviour, Radboud University Nijmegen

Paul Michel, Research Scientist, Google DeepMind

Roman Miletitch, Research Scientist, ThirdPlace Laboratory, Paris

Francis Mollica, Lecturer in Computational Cognitive Science, Complex Human Data Hub, School of Psychological Sciences, University of Melbourne

Juan Moriano, PhD candidate, University of Barcelona

Katie Mudd, Postdoctoral Researcher, Institute for Logic, Language, and Computation, University of Amsterdam

Thomas F. Müller, Research Scientist, Center for Humans and Machines, Max Planck Institute for Human Development, Berlin

Danielle Naegeli, PhD candidate, Department of Communication and Cognition, Tilburg University

Jonas Nölle, Postdoctoral Researcher, School of Psychology and Neuroscience, University of Glasgow

Xenia Ohmer, Postdoctoral Researcher, Natural Language Processing Group, Institute of Cognitive Science, Osnabrück University

Asli Özyürek, Professor and Director, Multimodal Language Department, Max Planck Institute for Psycholinguistics, Nijmegen, and Professor, Donders Institute for Brain, Cognition, and Behaviour, Radboud University Nijmegen

Aliki Papa, Postdoctoral Fellow, Department of Psychosocial Science, University of Bergen

David Peeters, Associate Professor, Department of Communication and Cognition, Tilburg University

Yakov Pichkar, Postdoctoral Researcher, Department of Biological Sciences and the Evolutionary Studies Initiative, Vanderbilt University

9780192886491_Book.indb 19 /12_first_proofs/first_pr…/xml_for_typesetting 27-Mar-25 23:02:15

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

AQ: "Olivier Pietquin, Research Director, Cohere" This is different from the description given by the author during the affiliation management, which is: 'Research Scientist, Google DeepMind'. Which is correct, or should they be combined?

Olivier Pietquin, Research Director, Cohere

Simone Pika, Professor, Comparative BioCognition, Institute of Cognitive Science, Osnabrück University

Ronald J. Planer, Lecturer, School of Liberal Arts, University of Wollongong, and Research Fellow, Words, Bones, Genes, and Tools, DFG Center for Advanced Studies, University of Tübingen

Michael Pleyer, Assistant Professor, Center for Language Evolution Studies, Nicolaus Copernicus University in Toruń

Andrea Ravignani, Professor, Department of Human Neurosciences, Faculty of Medicine and Dentistry, Sapienza University of Rome, Honorary Professor, Center for Music in the Brain, Department of Clinical Medicine, Aarhus University and the Royal Academic of Music Aarhus/Aalborg, and (former) Research Group Leader, Comparative Bioacoustics Group, Max Planck Institute for Psycholinguistics, Nijmegen

Limor Raviv, Research Group Leader, LEADS group, Max Planck Institute for Psycholinguistics, Nijmegen, and Lecturer in Social Interaction, cSCAN, University of Glasgow

Mathieu Rita, PhD candidate, INRIA

Seán G. Roberts, Lecturer, School of English, Communication, and Philosophy, Cardiff University

Wendy Sandler, Professor and Founding Director, Sign Language Research Lab, Department of English Language and Literature, University of Haifa

Marieke Schouwstra, Associate Professor, Institute for Logic, Language and Computation, University of Amsterdam

Anita Slonimska, Postdoctoral Researcher, Multimodal Language Department, Max Planck Institute for Psycholinguistics, Nijmegen

Rose Stamp, Lecturer, Department of English Literature and Linguistics, Bar-Ilan University

Florian Strub, Research Scientist, Cohere

Toshitaka N. Suzuki, Associate Professor, Research Center for Advanced Science and Technology, The University of Tokyo

Pui Yiu Szeto, Research Assistant Professor, Department of Asian and North African Studies, Ca' Foscari University of Venice

Shira Tal, Chancellor's Fellow, Centre for Language Evolution, University of Edinburgh

Monica Tamariz, Associate Professor, Psychology, Heriot-Watt University

9780192886491_Book.indb 20 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:15

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Claudio Tennie, Research Group Leader, Working Group Early Prehistory and Quaternary Ecology, Department of Geoscience, and member of Words, Bones, Genes, and Tools, DFG Center for Advanced Studies, University of Tübingen

Simon W. Townsend, Professor, Department of Evolutionary Anthropology and Center for the Interdisciplinary Study of Language Evolution, University of Zurich

Kristian Tylén, Professor, Department of Linguistics, Cognitive Science, and Semiotics, and Interacting Minds Centre, Aarhus University

Daria Valente, Researcher, Department of Life Sciences and Systems Biology, University of Turin

Carel P. van Schaik, Professor Emeritus, Department of Evolutionary Biology and Environmental Studies and Center for the Interdisciplinary Study of Language Evolution, University of Zurich

Stuart K. Watson, Postdoctoral Research Fellow, Department of Comparative Language Science, Department of Evolutionary Biology and Environmental Studies, and Center for the Interdisciplinary Study of Language Evolution, University of Zurich

Ray Wilkinson, Professor, Division of Human Communication Sciences, School of Allied Health Nursing and Midwifery, University of Sheffield

Isobel Wisher, Postdoctoral Researcher, Department of Linguistics, Cognitive Science and Semiotics, Department of Archaeology and Heritage Studies, and Interacting Minds Centre, Aarhus University

Noga Zaslavsky, Assistant Professor, Department of Psychology, New York University

9780192886491_Book.indb 21 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:15

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

9780192886491_Book.indb 22 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:15

Introduction

LIMOR RAVIV AND CEDRIC BOECKX

The evolution of language has been occupying the minds of philosophers and historians for over a millennium, thanks to our inherent interest in what makes our species 'special'—with language ranking high on everyone's list of traits contributing to this uniqueness. But understanding how different aspects of human language evolved is a major challenge: there are no archaeological records of the brains and languages of early humans who lived hundreds of thousands of years ago, and therefore we do not know what the first language(s) looked like or how exactly they were formed. Without a time-travel machine, it seems that questions surrounding the evolution of language can never be answered directly. And so for decades, linguists and cognitive scientists had essentially abandoned (and even called to ban) the study of language evolution. Alternatively, some scholars engaged in 'armchair linguistics', pondering and speculating about how language might have come about. Despite the absence of concrete data, such theories contributed an abundance of original ideas and were highly influential in shaping the field of language evolution (with many such hypotheses being put to the test today).

Thanks to the rise of modern technologies and novel methods that finally allow scientists to explore the ancient past and unpack the origin of human language, the last two decades have introduced a major data-driven paradigm shift in the field of language evolution, allowing it to move beyond speculations into the realm of the scientific method. Over the past twenty years or so, the field of language evolution has thus become a major area of empirical research, owing to genuine progress made within the community by developing new methods and tools that allow the empirical testing of prominent theories, as well as a gradual opening up of the field to foster interdisciplinary dialogue—borrowing and synthesizing helpful methodologies from related disciplines. This progress is well represented at major conferences such as Evolang and Protolang, but also by the increasing presence of articles on language evolution in prestigious journals whose scopes go well beyond the traditional boundaries of evolutionary linguistics.

Precisely because of this progress, it has become increasingly difficult for newcomers to keep track of new developments and tools in the field of language evolution, and particularly seeing as these advances increasingly rely on skill sets that are not (yet) part of the canonical curricula in Linguistics programmes in many universities around the

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

world. The overall aim of this Handbook is thus to help newcomers (as well as seasoned researchers) to appreciate and make sense of all these new developments, to provide a comprehensive map of the current landscape in the field (bearing in mind that it is a fast moving endeavour, and so new developments are likely to accumulate and complement the methods surveyed here), and to showcase some of the most important advances in the field—stressing that, contrary to a still widespread belief, it is in fact an empirical research field.

Our book is unique in putting emphasis on methodological aspects and the specificities of the different tools used in the field, with the goal of assisting readers in figuring out which of the many approaches that have emerged in recent years would be best suited for addressing specific questions. Although varied in style, each chapter is intended to: (i) highlight a particular methodology; (ii) articulate a question, or set of questions, that can be addressed using this methodology; and (iii) illustrate this methodology with at least one key example from the recent literature (typically, though not necessarily, taken from the work of the chapter's author(s)). Methodological limitations are discussed, and complementary approaches are featured as companion chapters. Nevertheless, our Handbook is definitely not a textbook: while it points to relevant methodologies, it does not incorporate training information and exercises to gain expertise. As such, it should serve as a guide to further, in-depth study. While readers will not become experts by reading the chapters alone, at the very least they will know what literature they have to get familiar with if they want to pursue their questions, as well as better understand the pros, cons, and basic procedures associated with each method.

The overall message of the volume is that multiple co-constraining approaches are necessary to do justice to the complexity of the problem of language evolution. Our initial objective was indeed to highlight the many fields of research that language evolution draws upon, from genetics to neuroscience, to artificial intelligence, computer science, experimental and comparative psychology, cognitive science, and linguistics.

Crucially, our volume does not intend to cover all the hypotheses and evolutionary scenarios considered about language evolution. Although more than a decade old, Oxford University Press's own *The Oxford Handbook of Language Evolution*, and Tecumseh Fitch's textbook *The Evolution of Language*, published in 2010, are still very valuable resources in the field in this regard. While we aimed to make our volume as comprehensive as possible, it is by no means an exhaustive list. We hope, though, that it is sufficiently representative. Furthermore, readers will not find chapters devoted to the (once dominant) generativist/minimalist tradition, given that we focus here on strictly empirical methodologies and data-driven approaches. Finally, we do not make a sharp distinction between language evolution and language change. In the past, such a distinction was often taken for granted: with language evolution research typically concerned with the transition from a state of no-language, 'protolanguage', or animal-like communication systems into a state of full-blown natural language with productive grammatical rules and expressive lexicons, while research on language change typically concerned the diachronic changes occurring in existing languages with established

9780192886491_Book.indb 24 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:15

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

grammars and vocabulary over the course of human history. However, under the now commonly accepted cultural evolution approach to language, which views language as a complex adaptive system that is adapted to fit its users' learning and communicative biases and needs, it seems that there are fundamental similarities in the processes and mechanisms underlying both scenarios.

We have decided to divide the Handbook into three main parts: the first focusing on human data; the second focusing on (computer) simulations; and the third on animal models.

- Part I on human data showcases the many ways in which research on contemporary humans, endowed with a capacity for language acquisition and use, can nevertheless shed light on the process of language creation/emergence/evolution when placed in specific experimental settings (e.g. when asked to learn or create artificial languages in the lab). This part of the Handbook also contains discussions of a variety of approaches exploiting human data, e.g. work on emerging sign languages, clinical studies testing the genetic underpinnings of language, or historical data such as ancient DNA and diachronic corpora, that can shed light on the deep history of our capacity for language.
- Part II is broadly devoted to simulations and computational models that enable the careful control of agents' environments, the biases ascribed to learners prior to the task, and the scaling up/reproducibility of certain tasks, seeing as runs on a computer can be repeated millions of times, and can involve a much higher number of learners than experiments done on humans in the lab. Here, we also highlight new and exciting technologies such as deep learning neural networks and swarm robotics, which are becoming increasingly popular in our field as in others.
- Part III revolves around the idea that non-human animals have a lot to tell us about language evolution, even though we may not often think of them as 'linguistic creatures'. It is now generally agreed upon that human language draws on many neurocognitive subtraits that are not specific/unique to humans. Identifying these in a wide range of species allows researchers to test specific questions that would not be available or ethical with human subjects. In addition, it explores comparative and observational work on non-human animals.

Readers should note that this tripartite division is a flexible one, and indeed, many approaches represented in one section complement those in another section. For instance, iterated learning models and communication games are represented in both Part I and Part II, because although they are different in their implementation, both experimental and computational approaches in these domains follow similar principles. Likewise, inferences based on ancient human genomes (discussed in Part I) often resort to insights from evolutionary neuroscience and comparative approaches looking for convergence and (deep) homologies across possibly quite distant taxa (Part III). Similarly, a lot of animal experiments also compare behaviour to human subjects, as is evident from several chapters in Part III. We thus encourage readers to seek connections

9780192886491_Book.indb 25 /12_first_proofs/first_pr/xml_for_typesetting 27-Mar-25 23:02:15

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

and synergies across different methods contained in a specific section. For example, we have explicitly commissioned multiple chapters on various approaches to gesture and sign languages (Part I) as well as on the syntactic organization of animal signals (Part III).

Our overall objective in putting this volume together was to stress the numerous empirical options available to tackle some of the most exciting questions in the field of language evolution. Obviously, these different approaches tap into distinct scientific cultures and practices. Newcomers should be aware that although we are keen to emphasize the complementarity of all these approaches, they do not yet have an equal representation at conferences or publication venues. Collaborating across fields is quickly becoming a must, but although similar research questions constitute a good initial common ground, the necessary training varies substantially, and it is important to note that not all methods are equally accessible. For example, while some of the approaches discussed here can be mastered within a relatively short amount of time and given modest equipment (e.g. artificial language learning experiments), others (e.g. comparative neuroimaging and genetics) may require extensive training periods, expensive resources, and larger teams of collaborators. The time required to collect data also varies considerably: from days to months, to years. While all of this is true, we are impressed by the progress made in the field as it took its empirical turn, and we are confident that the next decade will continue to remove disciplinary barriers and further enhance collaborative efforts.

Acknowledgements

We thank all authors for their willingness to contribute to this enterprise, and for their patience as we were coordinating schedules and deadlines. Many of the authors also helped us in reviewing the various submissions. In this context, we also want to express our gratitude to Pedro Tiago Martins, Marieke Woensdregt, Rie Asano, and Lukas Galke for reviewing some chapters. Thank you also to reviewers of our initial book proposal, for alerting us to potential pitfalls and highlighting a few gaps we hope we have managed to correct, and to Julia Steer at Oxford University Press for creating a home for this project.

9780192886491_Book.indb 26 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:15

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

PART I

HUMAN DATA

9780192886491_Book.indb 1 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:15

CHAPTER 1

ITERATED LEARNING EXPERIMENTS

MONICA TAMARIZ AND ALIKI PAPA

1.1 General principles of iterated learning experiments with a focus on language

The study of language evolution addresses two broad sets of questions. The first one explores the biological evolution of a new cognitive skill—language—in hominin species. The second set explores how cultural processes such as learning and usage shape linguistic structure. Iterated learning experiments address the second set of questions by modelling language learning, communicative interaction, and population dynamics using human participants and artificial miniature languages in the lab. The rest of this section provides a historical and disciplinary background for iterated learning experiments. The following sections (1.2–1.4) describe the most common methods employed to conduct iterated learning experiments and analyse the results. Section 1.5 discusses what this method contributes to cultural evolutionary theory, and the final section suggests some research questions that may be addressed in the future using this paradigm.

The general principle of iterated learning experiments involves training learners with some data and subsequently asking them to produce the same data. The input data to train each participant are the outputs produced by the previous participant (Figure 1.1). In this simple iterated learning chain (also called transmission or diffusion chain) design, each participant is a 'generation.'[1]

[1] 'Closed-group' studies, in which information is transmitted among members of a group, are covered in Tal et al. (Chapter 2, this volume).

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

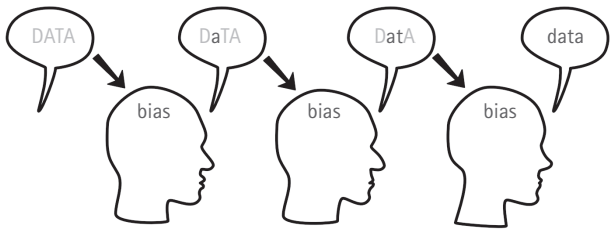


FIGURE 1.1 In this example, iterated learning chain, the outputs change in response to cognitive biases.

Recent iterated learning studies were foreshadowed by Esper's (1966) transmission-chain exploration of phonological change and Bartlett's (1932) influential 'reproduction chains' of drawings and stories—an experimental formalization of the 'Chinese whispers' or 'telephone' game.

1.1.1 Iterated learning helps reveal pressures on cultural evolution

A key tenet of iterated learning studies is that, after multiple iterations, the information that is transmitted changes and it adapts to factors that affect transmission such as learner's inductive biases (Griffiths et al. 2008), among other factors. The reason for this is that innovations and transmission errors introduced by participants are affected by those factors. For example, under the least-effort principle, errors tend to make the output simpler than the input. If members of a population share the same biases, effects accumulate over repeated learning and production events as the outputs evolve to match the biases (Figure 1.1). Iterated learning thus amplifies and reveals the effects of even very weak factors (Thompson et al. 2016). By examining the properties that emerge consistently over transmission chains, we can infer the forces that shape culture.

1.1.2 Iterated learning studies in humans

Many iterated learning experimental designs have sought to understand the cognitive, social, and cultural mechanisms of human cultural evolution. Studies have examined the change over experimental generations of different types of information. Drawings (Bartlett 1932; Tamariz and Kirby 2015), graphical patterns (Kempe et al. 2015), and stories and arguments (Mesoudi et al. 2006) change under the effects of pre-existing 'schemas' in the learners' minds. Sequences of colours (Cornish et al. 2013), consonants (Cornish et al. 2017), and musical and rhythmic patterns (Jacoby and McDermott 2017;

9780192886491_Book.indb 4 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:16

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Verhoef and Ravignani 2021; Lumaca and Baggio 2017; Miton et al. 2020; Angladа-Tort et al. 2023) evolve to match universal and local properties. And quantitative relationships (Kalish et al. 2007) become simpler, and therefore easier to learn or produce. These designs all involve reproduction tasks, i.e. participants were exposed to a drawing, a tune, etc., and were asked to reproduce it.

Other studies have used two-dimensional, form–function information: participants were asked to produce outputs to achieve a function. Fay et al. (2018) modelled the evolution of instructions to navigate a map and found the instructions became more effective over generations. Methods to solve technological problems also increased their effectiveness, e.g. Caldwell and Millen's (2008) spaghetti towers became taller and Derex et al.'s (2019) wheel task was solved more effectively after several iterations. Iterated *language* learning studies, which model the cultural evolution of signals that refer to meanings, fall within this category and are the focus of this chapter.

In a prototypical iterated language learning experimental design, participants are trained on a miniature artificial language—a set of linguistic signals paired with meanings—and are subsequently tested—presented with each meaning and asked to produce its associated signal. The output language produced by each participant is used as the input to train the next participant in a transmission chain.

1.2 Stimuli: Miniature artificial languages

Miniature artificial languages consist of a set of meanings, each paired with one signal (also called a label). The first participant in an iterated learning chain is trained with a 'seed' language.

1.2.1 Meanings

Countless possible sets of meanings, or *meaning spaces*, can be used to create languages. Figure 1.2 shows some examples from the literature. Depending on its internal structure, a meaning space can be structured, if it has clear dimensions and categories, or unstructured, if it does not have them. All the meaning spaces in Figure 1.2 (except 1.2f) are structured. Figure 1.2a has three dimensions: shade, shape, and motion, and each dimension includes three categories, e.g. for dimension 'shade', categories are 'black, 'grey', and 'white'. Meanings that have continuous properties (such as the triangles in Figure 1.2f) do not have *a priori* categories.

9780192886491_Book.indb 5 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:16

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

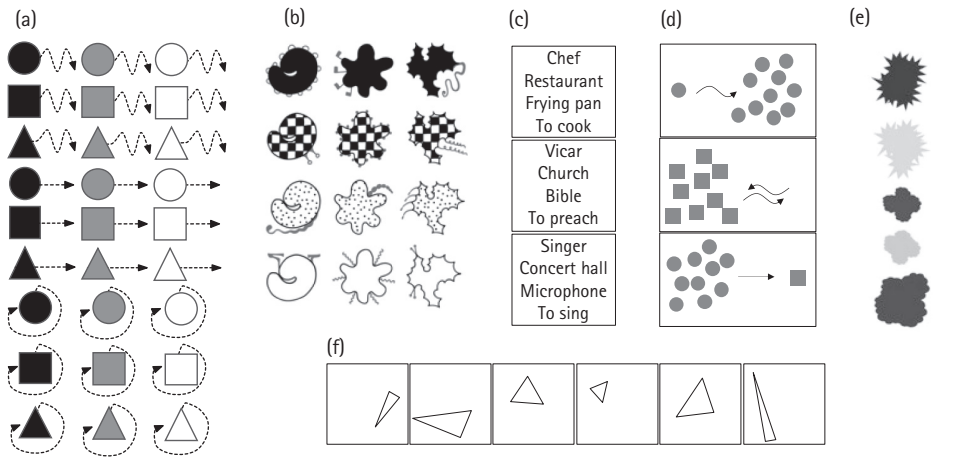


FIGURE 1.2 Example meaning spaces from iterated learning experiments.
A, B show the full meaning spaces used; the rest show subsets.

Sources: A: Kirby et al. (2008); B: Kirby et al. (2015); C: Motamedi et al. (2019); D: Saldana et al. (2019); E: Carr et al. (2017); F: Kempe et al. (2019).

Interesting meaning-related effects in the literature include the finding that categories may emerge, for instance, in a triangle meaning space (Figure 1.2f) over iterated learning (Carr et al. 2017). Perfors and Navarro (2014) and Silvey et al. (2015) show that language structure evolves to reflect the salience of different meaning dimensions. The colour categories that emerged in Müller et al.'s (2021) long iterated language game chains were similar to the colour categories in the participant's native languages. Verhoef et al. (2016) examined how spatial graphical signals evolved to refer to temporal meanings. And Ma and Fiveash (2019), whose meaning space included two categories of meanings—emotional and referential—found that signals for the former became music-like and, for the latter, speech-like.

1.2.2 Signals

The signals are often written pseudowords or phrases that participants type or speak. The signals in the languages shown in Figure 1.3 were random concatenations of two, three, or four syllables sampled from a set of nine.

Signals in iterated language learning experiments are not always world-like. They can be musical, such as slide-whistle tunes (Verhoef 2012; Verhoef et al. 2014) or combinations of two tones (Kempe et al. 2019); spoken signals (Smith and Dediu 2020); drawings (Theisen-White et al. 2011; Caldwell and Smith 2012); gestural signals (Motamedi et al. 2019; see also Naegeli and Schouwstra (Chapter 9, this volume)); a set of figurative—e.g. a balloon, a pencil, etc.—and abstract symbols (Müller et al. 2021); and even Lego constructions (Tamariz et al. 2016). Further modalities, such as haptic patterns played on the skin, dance signals, knitting patterns, etc., may be used in the future to explore the affordances of different communication channels.

9780192886491_Book.indb 6 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:17

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

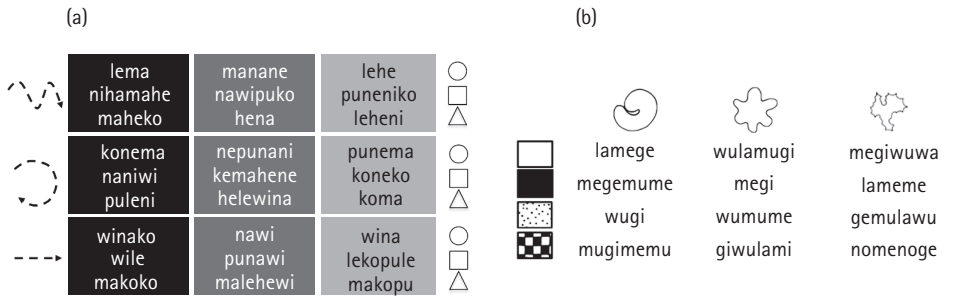


FIGURE 1.3 Two seed languages.

The position of the signal in each figure indicates the combination of dimensions of the meaning it refers to (e.g. on the top left, 'lema' is the signal for a black circle with a bouncy movement arrow).

Sources: A: Kirby, Cornish, and Smith (2008); B: Kirby, Tamariz, et al. (2015).

The signals for seed languages are not always determined by experimenters; they may be generated by participants, as in the case of Theisen-White et al.'s (2011) drawings or Motamedi et al.'s (2019) gestures, which were produced by participants when prompted by meanings like those in Figure 1.2c.

1.2.3 Mappings

There are different types of mappings between meanings and signals: in *compositional* languages (e.g. Figure 1.4a), the signal for a complex meaning (e.g. light grey bouncing

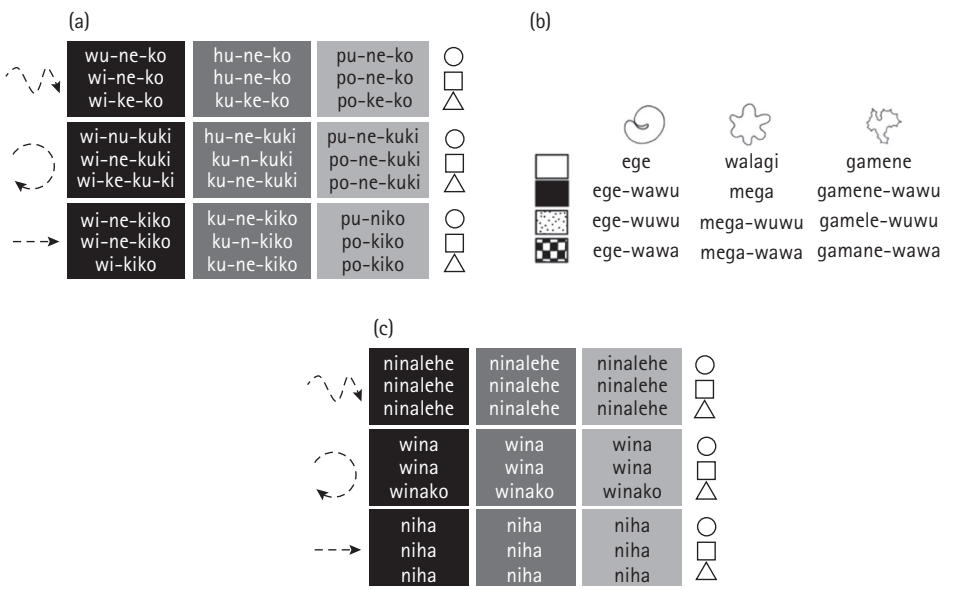


FIGURE 1.4 Three output languages produced at the last generation of iterated learning chains.

9780192886491_Book.indb 7 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:17

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

ball) is a function of the meanings of its elements (po = light grey; ko = bouncing) and the way they are combined (shade—first segment; motion—last segment). The language in Figure 1.4b is fully expressive, as it can clearly express each meaning distinctly. It is almost perfectly compositional and contains only a few irregularities. Compositionality is found in natural languages in morphosyntax and phonology and, at the lexical level, in regular systems such as numerals or the names of playing cards (Tamariz 2011).

The position of each signal in the tables indicates the combination of meaning dimensions of their associated meaning. A (based on Kirby et al. 2008): Experiment 2, in which participants were trained on 50 per cent of the items selected from the previous participant's output language and had to produce signals for 100 per cent of the meanings. The iterated learning chains in this experiment had an *anti-homonymy filter* designed to generate an artificial pressure for expressivity: the training items were selected to avoid repeated signals as far as possible. B (based on Kirby et al. 2015): a communicative task was embedded within each generation of the iterated learning chain (see Figure 1.4). C (based on Kirby et al. 2008): Experiment 1, in which participants were trained on 50 per cent of the items selected *randomly* from the previous participant's output language and had to produce signals for 100 per cent of the meanings. This language is systematic, but not compositional, and has low expressivity, as it does not differentiate between shapes.

In *systematic* languages, similar signals refer to similar meanings (Figure 1.4c). All compositional languages are also systematic (e.g. in Figure 1.4b, all meanings with the first shape start with 'ege'). But not all systematic languages are compositional (Figure 1.5).

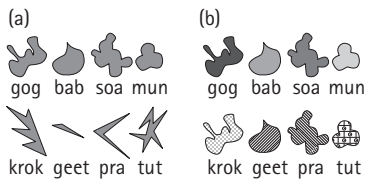


FIGURE 1.5 Two languages that exhibit systematicity but not compositionality.

A: Spiky shapes have plosive consonants, round shapes do not. B: Shapes with crosshatching have plosive consonants, shapes with block colour do not. Language A has the additional property of iconicity, whereby signals intuitively resemble, or are motivated by, the meanings.

Holistic languages have a distinct signal for each complex meaning. An example would be having a different proper name (Sam, Alex …) for each card in the deck. *Homonymy* is a relevant property of language, with the same signal for more than one meaning, which makes languages more learnable but introduces ambiguity thus reducing communicative or expressive power. Languages that feature homonymy are often called *degenerate* (Kirby et al. 2015) or *underspecified* (Smith 2018) in the iterated learning literature. Other properties of the signal–meaning mappings relevant to iterated learning studies include hierarchical structure (e.g. Saldana et al. 2019), iconicity (the resemblance between form and meaning present for instance in the language

9780192886491_Book.indb 8 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:18

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

in Figure 1.5a), and arbitrariness (the opposite property, whereby the link between forms and meanings is unmotivated; for instance, in most iterated learning chains, different signals evolve for the same meaning *across chains* almost invariably).

1.3 Procedure for iterated language learning studies

1.3.1 Iteration and population structure

In iterated *linear chains* (Figures 1.1 and 1.6A), there are one or more participants in each generation. They are trained on an input language (section 1.3.2) and later tested (section 1.3.3), and the output they produce in the test is then used to train the next participant(s) in the chain. Linear chains provide a simple model of generational turnover and help isolate the contribution of learning by fresh minds to linguistic structure. In *replacement chains* (Figure 1.6B), a group of participants interact; at each generation, one or more participants leave the group and new ones (in light grey in Figure 1.6B) join the group. In this type of chain, participants sometimes enter the group directly; alternatively, they may be trained before joining. It is common for newcomers to observe the group interactions for some time before they start actively taking part. Replacement chains model a gradual generational turnover and include some interaction among individuals with different degrees of expertise.

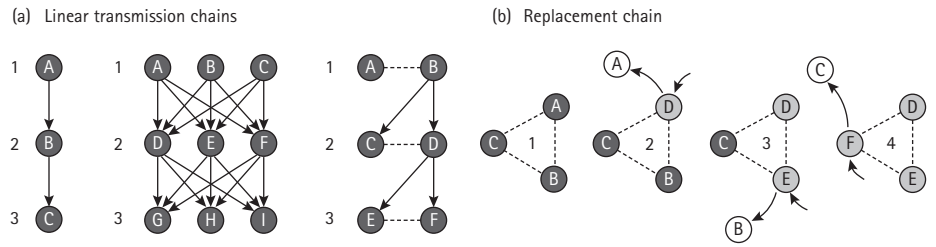


FIGURE 1.6 Population structure in iterated learning chains.

Numbers = generations; letters = individuals; solid arrows = transmission; dashed arrows = interaction. a: Linear chains. Left: a pure transmission chain with one participant per generation; Middle: another one, with three participants per generation. Right: a chain with two participants per generation including transmission and interaction (e.g. communication). b: A replacement chain. By the end of this chain, all the individuals in the initial generation (or round) have been replaced by fresh ones.

1.3.2 Training

During training, participants typically see a meaning with its accompanying signal on a computer screen for a few seconds, followed by a brief blank screen before the next

9780192886491_Book.indb 9 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:18

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

meaning and signal appear (Figure 1.7(1)). Training duration and regime can vary considerably across studies. The training block in Kirby et al. (2015) included six rounds of exposure to the full language. Items were presented in random order within each round. By contrast, Kirby et al.'s (2008) study applied a *transmission bottleneck*, i.e. each participant was trained on only a subset of the complete language, but during testing, they were asked to produce signals for all the meanings. Figure 1.4c is an output language from this study. Bottlenecks can be used as a model of imperfect learning and to test whether the language is capable of generalization to novel, unseen items. They can also be used to introduce selection pressures on evolution. In the second experiment in Kirby et al. (2008), for instance, care was taken to avoid duplicate signals in the selection of items for the next participant's training language (Figure 1.4a). This resulted in languages that had few homonyms, and therefore could be more expressive, as distinct signals could be associated with each meaning. This pressure from expressivity, together with the ubiquitous pressure for learnability, resulted in compositional languages.

The order of presentation of items during training is typically random, but it can be scaffolded, e.g. items can be presented incrementally, with a reduced number of items or meaning dimensions shown a few times before new items or dimensions are gradually introduced. This manipulation usually facilitates learning. To achieve a more active form of learning, training trials may be interspersed with test trials in which participants observe and produce, rather than simply observe, from the early stages of the study. This allows, additionally, to train participants to criteria, that is, to give them successive training and testing blocks until they respond correctly to a pre-established proportion of trials in a test block. Kempe et al. (2021) used this method to manipulate the amount of learning given to participants and to measure the learnability of languages. Training to criteria filters out variation in participants' outputs due to individual differences in attention span, concentration, etc., from the effects of transmission, which tend to be the focus of iterated learning studies.

A more active and naturalistic training regime is composed of 'comprehension trials' only (similar to the task in Figure 1.7(2a)). In these trials, participants are asked to select the correct meaning for a given signal from an array, and are subsequently given partial or full feedback, which may include whether their selection was correct or incorrect and, if incorrect, pointing at the correct meaning. Following a training block, participants can then be tested in blocks of production trials (similar to the task in Figure 1.7(2b)), in which they are shown one meaning at a time and asked to produce the corresponding signal, discretionarily followed by feedback (Kempe et al. 2021).

1.3.3 Testing

Testing involves displaying meanings in turn to the participant and asking them to produce (type, speak, sign, etc.) the signal associated with each of them.

Testing trials can include communicative tasks to make them more ecologically valid. See an example in Figure 1.7. In these tasks, pairs of participants learn and interact

9780192886491_Book.indb 10 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:18

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

communicatively together, either in person (Theisen-White et al. 2011) or through computer monitors (Kirby et al. 2015; Figure 1.7).

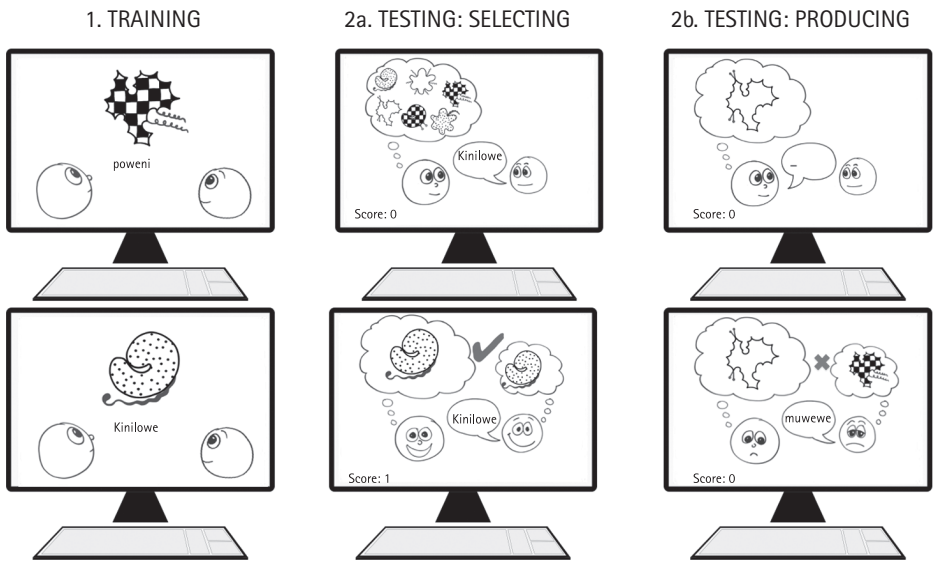


FIGURE 1.7 Screen captures from the procedure in Kirby et al.'s (2015) study, which included a collaborative, communicative testing task.

Participants interacted in pairs. Each one was instructed to imagine she was the character on the left throughout, while their partner was the character on the right. 1: A training trial, in which the participants saw each item in the language in turn. 2: A test trial seen from the two participants' points of view. 2a: A test trial in which the partner produced one of the signals, and the participant selected from a subset of the meaning space. In this case, the participant selected the correct item (the item selected matched the signal) and the score increases (note, this is similar to the comprehension *training* trials in Kempe et al. (2021), described in the previous section, but here they are inserted in a communicative task). 2b: A test trial in which the participant was prompted with a meaning and had to type the signal for their partner to guess that meaning. A point was not scored because the item guessed did not match the prompt.

Müller et al. (2021) used comprehension (selection) tests in their large-scale online study. Since they were investigating the emergence of categories, or new mappings between symbols and colours, they gave no feedback regarding whether the choices were correct or not. The mappings crystallized over continued interaction.

1.3.4 Choices in iterated learning experimental design

Many of the features of iterated learning experiments are shown in Figure 1.8. This visualization can help inspire new hypotheses and generate new designs for future studies.

9780192886491_Book.indb 11 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:19

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

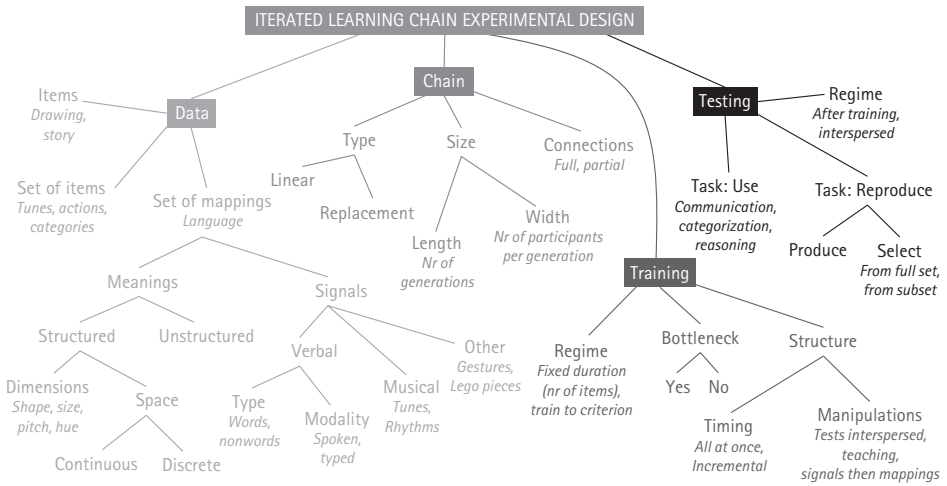


FIGURE 1.8 Some choices in transmission-chain iterated learning experimental design. Examples and explanations in italics.

1.4 Analysing the outputs of iterated language learning chains

The analyses used in iterated learning experiments are designed to quantify the evolution of the outputs produced by participants over generations. Properties of the outputs can evolve as adaptations to faithful transmission (i.e. outputs become simpler, easier to learn, and reproduce) and/or to other functions such as expressivity.

1.4.1 Transmission fidelity

Transmission fidelity, also referred to as learnability (or its opposite, transmission error), is measured as the similarity (or its opposite, difference, also referred to as distance) between one participant's input and their output. Similarity (or distance) metrics include Hausdorff distance (Huttenlocher et al. 1993) for simple drawings; (normalized) Levenshtein edit distance (Levenshtein 1966) for strings (of letters, characters, sounds, colours, etc.); Hamming distance (Hamming 1950) for items defined by a finite sets of properties; and subjective similarity ratings collected from naive participants, which are used when similarity or distance are difficult to measure automatically (e.g. Carr et al. 2017).

1.4.2 Structure

For our purposes, structure is defined as departure from randomness and compressibility. Structured outputs are more predictable, more regular, simpler, and usually easier

9780192886491_Book.indb 12 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:19

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

to learn than random ones. We can measure structure at different levels. For individual items, metrics will be specific to the type of data. For example, structure in drawings can be quantified with Perimetric complexity (Pelli et al. 2006) or Kolmogorov (or algorithmic) complexity (e.g. Tamariz and Kirby 2015; Kempe et al. 2021); in stories, with the number of word types, word type–token ratio in the language (Berdicevskis and Semenuks 2022) or with different measures of syntax complexity.

For sets of items, for example for a set of signals, structure can take the form of dimensions shared by several items, which define categories. Combinatoriality quantifies the extent to which sub-elements are reused across items and can be measured using Form Recombination Index (FRI) (Galantucci et al. 2010) or the entropy of the segments in the items in the set (Verhoef et al. 2014; Motamedi et al. 2019). Hierarchical structure, or the presence of nested patterns, can be estimated in a set of strings (of characters, musical notes, colours, etc.) using Associative Chunk Strength (Knowlton and Squire 1994; Cornish et al. 2017).

When it comes to measuring structure in the signal–meaning mappings of miniature artificial languages, the following sections detail some commonly used metrics of systematicity and compositionality.

1.4.2.1 Systematicity

Systematicity, or the degree to which signals that are similar refer to meanings that are also similar, has been quantified using the Mantel test (Tamariz 2008; Kirby et al. 2008, 2015; among many others). In this Monte-Carlo type test, first we obtain the similarity (or distance) between every pair of signals and between every pair of meanings and then we test to what extent the correlation between the two pairwise–similarity lists is significant (Figure 1.9).

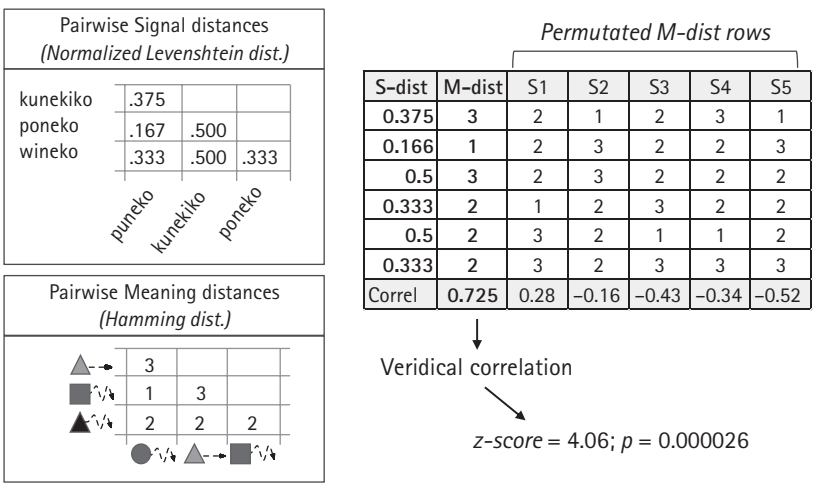


FIGURE 1.9 Mantel test of systematicity in a subset of signals and meanings in the output language from Kirby et al. (2008) shown in Figure 1.4b.

Left: Distances between pairs of signals (top) and pairs of meanings (bottom). Right: Distances copied into the first two columns of the table. In the remaining columns are scrambled versions of the meaning distances (obtained by permutating the rows). The bottom row shows the correlations between the signal pairwise distances and each of the other columns (in a larger set of signals and meanings, we would obtain a large number of random permutations, e.g. 5,000 or 10,000 for robustness of the statistical test). The *z*-score of the veridical correlation in the distribution of all correlations and the corresponding *p* value indicates how likely it is that the veridical correlation occurred by chance.

9780192886491_Book.indb 13 /12_first_proofs/first_pr…/xml_for_typesetting 27-Mar-25 23:02:20

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Beckner et al. (2017) reanalysed Kirby et al.'s (2008) data using an alternative method to estimate the systematicity of each meaning dimension. They took all the signals referring to a meaning (e.g. colour), containing a category (e.g. 'red'), calculated the normalized Levenshtein edit distance for all the possible signal pairs, and subtracted it from 1. Averaging the result over all colours approximates how systematically 'colour' is encoded in the language.

1.4.2.2 Compositionality

The regularity of the mappings between components of signals and meaning categories is captured by *RegMap* (Tamariz and Smith 2008; Cornish et al. 2009), an information-theoretic metric based on conditional entropy that returns an overall compositionality value for a language (Equation 1) and partial regularities between individual features of meanings and signals and separate values for comprehension and production. *RegMap* requires segmenting signals and meanings into features such as properties (size, shape, etc.) or elements (syllables, letters, sounds, etc.). The calculation of *RegMap* is summarized in Figure 1.10.

Equation 1

$$RegMap = \sqrt{\left(1 - \frac{H(S|M)}{\log(s)}\right) \times \left(1 - \frac{H(M|S)}{\log(n_m)}\right)}$$

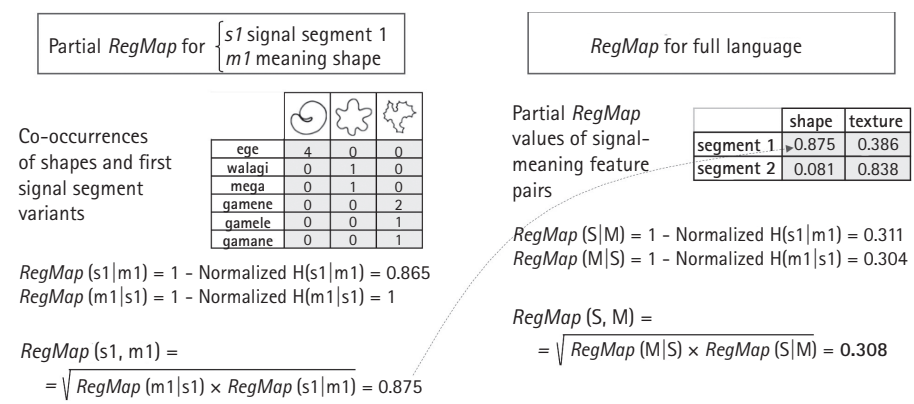


FIGURE 1.10 Calculation of *RegMap* of the last-generation language from Kirby et al. (2015) shown in Figure 1.4b.

Each signal was segmented into three segments: initial, middle, and final. Top: overall *RegMap* formula to estimate the compositionality of the language. Bottom left: example calculation of partial *RegMap* for one feature of the signals (initial segment) and one feature of the meanings (shape). *RegMap* (m1|s1) relates to comprehension and has a value of 1 because given any initial signal segment,

9780192886491_Book.indb 14 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:20

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

we can be completely certain of which meaning it refers to. *RegMap* (s1|m1) relates to production and, in this case, it is less than 1 because for the second and third shapes there are several possible initial segments. To obtain partial *RegMap* (s1, m1), we apply the formula shown on the very bottom left. Bottom right: to calculate *RegMap* for the full language, we obtain all partial *RegMaps* and use this matrix to obtain *RegMap* (M|S) for comprehension and *RegMap* (S|M) for production. Based on these two values, the overall *RegMap* (S, M) of the language is calculated using the formula at the very bottom, right. All *RegMap* values range between 0 and 1.

In the examples employed above to illustrate systematicity and compositionality, meaning features were designed to be very clear (colours, shapes …), but other meaning spaces may be continuous, or their features may not be obvious. This is the case in Carr et al.'s (2017) triangles (Figure 1.2f) or Xu et al.'s (2010) colour space. When categories emerge over iterated learning in these meaning spaces, structure can be quantified by examining the extent to which individual signals are consistently mapped onto a set of meanings that are similar in some respect. Of course, the same set of meanings may be categorized in different ways by different participants or across iterated learning chains. Once categories are in place, systematicity (and compositionality if several dimensions are present) may emerge and be measured using the methods above even in this type of continuous meaning space.

Other methods to quantify linguistic structure in the literature exist for hierarchical structure (Saldana et al. 2019); iconicity of signal–meaning mappings (Tamariz et al. 2018); and sublexical structure in a continuous meaning space (Carr et al. 2017). Properties that are difficult to measure objectively such as non-compositional systematicity or iconicity can be quantified by collecting ratings from participants.

1.4.3 Evolution

Changes in properties of the participants' outputs across generations (such as fidelity, systematicity, iconicity, compositionality, etc.) can be measured with metrics such as Page's L (Page 1963, used by e.g. Caldwell and Millen 2008; Wasielewski 2014), growth curve analysis, time series analyses, or regressions. Differences between adjacent generations (0 and 1, 1 and 2, 2 and 3, etc.) quantify the change introduced by each generation. Differences between the seed input and the output of each generation reveals increasing divergence from the origin, or cumulative evolution.

In complex meaning spaces, it is possible to train each generation with a different selection of meanings. To enable an analysis that compares like with like, in addition to the changing set of meanings to be passed on to the next generation in the transmission chain, participants can be tested on a fixed set that remains constant across generations.

9780192886491_Book.indb 15 /12_first_proofs/first_pr…/xml_for_typesetting 27-Mar-25 23:02:21

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

1.5 Contributions of iterated learning experiments to cultural evolutionary theory

1.5.1 The expressivity–simplicity trade-off

The iterated learning design allows us to measure a key mechanism of cultural evolution, which different studies characterize as emergence, adaptation, or cumulative cultural evolution (CCE). When a set of discrete items such as signals evolve, we can use population thinking and define evolution as 'change in proportions of variants in a population'. If, additionally, evolution results in the signal population evolving *in response to*, or *to match* something in the items' environment, such as inductive biases, the structure of meanings, an additional pressure (e.g. for communication or coordination), or other constraints on transmission (e.g. limited access), we can speak of adaptation.

The evolution of linguistic structure can be represented in a space defined by two main dimensions—simplicity and expressivity (Figure 1.11; Kirby et al. 2015; see also Kemp and Regier 2012): *simplicity* is related to ease of learning, efficiency, and least effort. Languages evolve to be simple under constraints on (or as an adaptation to) transmission and usage factors such as the introduction of fresh learners (Kirby et al. 2008), an expanding meaning space (Raviv et al. 2019), or limited resources and opportunities to learn. A reduction of information (e.g. a degenerate language in which all utterances end up being 'a') or information that is compressible (any structure, any deviation from randomness) increases simplicity values. Simplicity and ease of learning, in turn, support transmission fidelity.

Expressivity is related to the degree to which utterances for different meanings are distinct and unambiguous; it supports the communicative function of language. Languages maintain or develop expressivity to the extent that it is needed for communication. For example, if a distinction between two meanings is never required during communicative interactions, the difference between the signals for those meanings may never emerge or may be lost (Silvey et al. 2015; Winters et al. 2015). Alternatively, if a distinction that is required is very complex, the language may overspecify, i.e. provide redundant information for that distinction (Tinits et al. 2017).

Simplicity and expressivity have orthogonal 'interests' and guide language evolution in opposite directions. Given a set of resources (memory capacity, existing linguistic structures, meaning space, etc.), the set of possible languages in the simplicity–expressivity two-dimensional space is contained within a Pareto frontier (Figure 1.11).

9780192886491_Book.indb 16 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:21

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

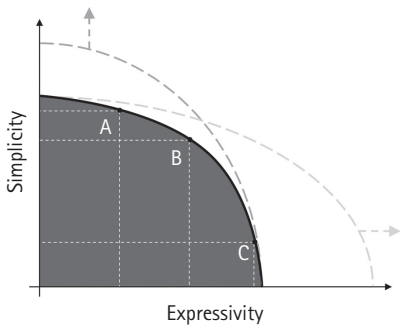


FIGURE 1.11 A Pareto frontier (the black line) contains the area of possible languages (shaded area). Languages that are on the frontier line optimize two desirable properties: simplicity and expressivity.

Language 'A' is very simple, but not very expressive (like the language in Figure 1.5a), while 'C' is very expressive, but not very simple (an example will be a language in which each meaning has a unique name: Sam, Alex, etc.). Language 'B' has high, but not maximal, levels of simplicity and expressivity. As they evolve, languages move within the shaded area. Under selection for expressivity and simplicity, they tend to do so towards the frontier, adapting to communication (rightwards) and/or to learning biases (upwards). Once at the frontier, they may move along it. If the affordances change, the frontier may move. If, say, new meanings enter language 'C', it may become more expressive and move to the left in the graph, bringing the frontier with it (right-pointing arrow). If, say, compositionality emerges in language 'A', the language moves up bringing the frontier with it (upwards-pointing arrow) (the frontier can also move 'back', leftwards of downwards if expressivity or simplicity are lost).

Iterated learning experiments allow us to manipulate those resources. These miniature models of cultural evolution amplify and reveal biases, and so they allow us to quantify adaptation to the selective pressures, as well as neutral evolution, in the output languages.

To illustrate this, consider the languages in Kirby et al. (2008, 2015). The seed languages taught to the first participant in each chain were holistic—each meaning was associated with a different, randomly constructed signal. They were fully expressive, but too complex to learn faithfully with the amount of training available. We can position them at point 'C' in Figure 1.11. Over generations, they gained simplicity. In the case of Kirby et al. (2008: experiment 1), when the languages were simply being learned and reproduced, they lost expressivity and gained simplicity by reducing the number of distinct signals (Figure 1.4c). They moved up and left, along the frontier. The languages in Kirby et al. (2015), which were learned over several generations and additionally used for communication (Figure 1.4b), and the languages in Kirby et al. (2008: experiment 2), under anti-homonymy filtering (Figure 1.4a), became compositional, which simultaneously maintained full expressivity and gained simplicity. They moved up in the space in Figure 1.11, which might seem like they leave the area of possible languages. Instead, the emergence of compositionality, by changing the two-dimensional (2D) landscape and its affordances, moves the frontier up.

1.5.2 The evolution of linguistic structure

How exactly does linguistic structure evolve in the face of expressivity and simplicity pressures? In our prototypical examples (Kirby et al. 2008, 2015), the seed languages

9780192886491_Book.indb 17 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:21

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

already possess *symbolicity*, as it is clear to participants that signals refer to meanings. Initially, *holistic* signals are unambiguously mapped onto meanings during training, making the languages fully expressive. Many signal elements are pruned during the initial generations. Later, *combinatoriality* emerges when the few signals left are combined in different ways. As dimensions and categories are inferred from the meanings, new, *finer-grained symbolic* associations are made between these on one hand and signal elements on the other (e.g. dimensions associated to position; categories associated to syllables), and *compositionality* emerges. Beyond these studies, other language properties such as *iconicity* (Tamariz et al. 2018), *predictability*, or conditional dependencies of grammatical markers (Smith and Wonnacott 2010), and *hierarchical structure* (Saldana et al. 2019) have emerged in iterated learning chains.

These studies provide an exciting model of cultural evolution and its mechanisms. We now illustrate *replication*, or faithful reproduction of elements of the input signals, with an analysis of the seed and final languages from Kirby et al. (2015), shown in Figures 1.3B and 1.4B, respectively. The seed language contained nine distinct syllables, of which six persisted until the final generation. The participants in generations one to six introduced, respectively, only two, three, one, zero, zero, and one syllables each. Moreover, of the ten letters in the seed language, nine survived until the last generation, and no other letters were added. *Innovation* happened when novel items such as syllables, and the removal, or pruning, and recombination of signal elements were introduced. Finally, *selection* involves the adaptation of the language to factors such as cognitive biases and meaning space structure, and is mirrored in the emergence of structure at the many levels described in the previous paragraph. Figure 1.12 illustrates the adaptation of signal segments to the functional niches in the meaning space, shape, and texture, over generations.

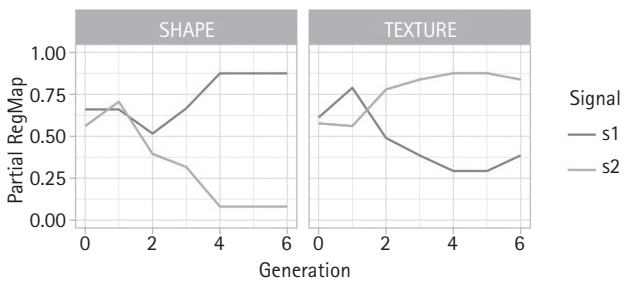


FIGURE 1.12 Regularity of the mappings (partial *RegMap*) between the two signal segments and two meaning dimensions over generations in a chain (based on Kirby et al. 2015).
Seed language is in Figure 1.3b; final language is in Figure 1.4b, analysis is in Figure 1.9.

1.5.3 Problems and benefits of iterated learning experiments

In an iterated language experiment, the unit of analysis is the chain, rather than the individual. The large numbers of participants required for robust statistical analyses may

9780192886491_Book.indb 18 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:21

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

present an access or funding problem. In addition, when participants from different generations have to be present in the same session, e.g. to construct physical artefacts (as in Caldwell and Millen 2008) or to enable subtle interpersonal interactions, this poses a practical problem. Coordinating participants can become extremely complex. An answer to this is to run tasks and interactions online, for example, embedded in general-purpose online platforms such as Mechanical Turk or academic-oriented ones such as Prolific, or built in purpose-made experiment management platforms that offer diffusion chain designs such as Dallinger (https://github.com/Dallinger/Dallinger). Online studies have the distinct advantage of allowing fast recruitment and turnover of large numbers of participants.

The use of iterated language learning studies as models of the origin and evolution of language has been criticized, notably because the participants are modern, adult, literate, enculturated humans, who are very different both from early humans at the time when language first evolved in our species, and from the infants who learn language in real life. Moreover, these studies take minutes or hours, much shorter than the years of individual language learning or the decades of generation turnover, and involve very limited numbers of speakers. Design features of iterated learning experiments circumvent some of these issues. Participants who are not familiar with signal modalities such as gesture or drawings, and novel meaning spaces, such as triangles or colours, can be found easily. A good model of illiteracy can be achieved by using non-musician participants in a study with musical signals (e.g. Tamariz et al.'s 2010 non-iterated study of the effects of literacy on structure). Comparing adult and child participants explores their distinct contributions to language change and evolution. Experiments in non-human animals explore the effects of biases we may share with primates, vocal learners, etc. Using miniature artificial languages and manipulating experimental population sizes, as well as combining lab experiments with computer simulations, helps us extrapolate to realistic timescales and population sizes.

Iterated learning experiments do, however, shed light on how languages adapt to the pressures posed by participants, meanings, and tasks. For example, they are an excellent model of the origin and evolution of emerging sign languages, which develop linguistic structure from scratch in communities over time through usage and transmission. Iterated learning studies thus contribute to cultural evolutionary theory by modelling the cultural evolution of language and, more generally, culture, as the result of biases on transmission and usage.

In revealing biases on transmission and their effects, these experiments shift part of the burden of explanation of cultural variation and structure away from innate capacities and towards cultural processes (Thompson et al. 2016; Kirby 2017). They also shift the focus of language and cultural origins away from the biological evolution of pre-conditions and towards the co-evolution of the human organism and the cultural niche it constructs.

9780192886491_Book.indb 19 /12_first_proofs/first_pr…/xml_for_typesetting 27-Mar-25 23:02:21

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

1.6 Research questions for the future

Iterated learning obtains *solutions that optimize* multiple constraints in complex tasks. Within language evolution, it can help further our understanding of multiple underexplored areas such as the *developmental course of learning biases*. We know little about the contribution of different age groups to linguistic structure (children, adolescents, young and older adults; or parents vs. non-parents), and iterated language learning studies with young children and infants, in particular, have proved challenging (but see Kempe et al.'s 2019 iterated language learning study with 4- and 5-year-olds). New experimental designs using, for instance, implicit testing methods, such as eye-tracking to select a meaning from an array, could be developed to advance in this direction.

The *affordances of new communication channels* can also be explored using iterated learning experiments. Studies using visual or haptic, tactile signals could inform future virtual reality applications, for instance.

Many studies have combined computer models with iterated learning experiments. This combination allows us to both *generate and test hypotheses* regarding the causes of language structure. Computer simulations with bias values informed by experiments can help us understand long-term and large-population scale effects (see Part II of this volume).

Psychological *experiments on learning*, such as effects on recall, priming, etc., can be iterated to amplify and test effects that may be too weak to become apparent after a single instance of learning.

Finally, language universal properties can (continue to) be explored by testing whether they emerge or are maintained in iterated learning chains including participants from many backgrounds (e.g. different first languages, different cultures) distributed either within the same chains or across different chains.

References

Anglada-Tort, M., Harrison, P. M., and Jacoby, N. (2023). 'Large-scale iterated singing experiments reveal oral transmission mechanisms underlying music evolution', *Current Biology* 33(8): 1472–1486, https://doi.org/10.1016/j.cub.2023.02.070.

Bartlett, F. C. (1932). *Remembering: A Study in Experimental and Social Psychology*. Cambridge: Cambridge University Press.

Beckner, C., Pierrehumbert, J. B., and Hay, J. (2017). 'The emergence of linguistic structure in an online iterated learning task', *Journal of Language Evolution* 2(2): 160–176, https://doi.org/10.1093/jole/lzx001.

Berdicevskis, A. and Semenuks, A. (2022). 'Imperfect language learning reduces morphological overspecification: Experimental evidence', *PLoS ONE* 17(1): e0262876, https://doi.org/10.1371/journal.pone.0262876.

Caldwell, C. A. and Millen, A. E. (2008). 'Experimental models for testing hypotheses about cumulative cultural evolution', *Evolution and Human Behavior* 29: 165–171.

9780192886491_Book.indb 20 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:21

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Caldwell, C. A. and Smith, K. (2012). 'Cultural evolution and perpetuation of arbitrary communicative conventions in experimental microsocieties', *PLoS ONE* 7(8): e43807, https://doi.org/10.1371/journal.pone.0043807.

Carr, J. W., Smith, K., Cornish, H., and Kirby, S. (2017). 'The cultural evolution of structured languages in an open-ended, continuous world', *Cognitive Science* 41(4): 892–923.

Cornish, H., Tamariz, M., and Kirby, S. (2009). 'Complex adaptive systems and the origins of adaptive structure: What experiments can tell us'. Special issue on 'Language as a complex Adaptive System', *Language Learning* 59: 4S1.

Cornish, H., Smith, K., and Kirby, S. (2013). 'Systems from sequences: An iterated learning account of the emergence of systematic structure in a non-linguistic task', *Proceedings of the Annual Meeting of the Cognitive Science Society, UC*. Merced.

Cornish, H., Dale, R., Kirby, S., and Christiansen, M. H. (2017). 'Sequence memory constraints give rise to language-like structure through iterated learning', *PLoS ONE* 12(1): e0168532, https://doi.org/10.1371/journal.pone.0168532.

Derex, M., Bonnefon, J. F., Boyd, R., and Mesoudi, A. (2019). 'Causal understanding is not necessary for the improvement of culturally evolving technology', *Nature Human Behaviour* 3: 446–452, https://doi.org/10.1038/s41562-019-0567-9.

Esper, E. A. (1966). 'Social transmission of an artificial language', *Language* 42(3): 575–580, https://www.jstor.org/stable/411408.

Fay, N., Ellison, T. M., Tylén, K., et al. (2018). 'Applying the cultural ratchet to a social artefact: The cumulative cultural evolution of a language game', *Evolution and Human Behavior* 39(3): 300–309, https://doi.org/10.1016/j.evolhumbehav.2018.02.002.

Galantucci, B., Kroos, C., and Rhodes, T. (2010). 'The effects of rapidity of fading on communication systems', *Interaction Studies* 11(1): 100–111.

Griffiths, T. L., Kalish, M. L., and Lewandowsky, S. (2008). 'Theoretical and empirical evidence for the impact of inductive biases on cultural evolution', *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 12* 363(1509): 3503–3514, https://doi.org/10.1098/rstb.2008.0146.

Hamming, R. W. (1950). 'Error detecting and error correcting codes', *Bell System Technical Journal* 29(2): 147–160.

Huttenlocher, D. P., Klanderman, G. A., and Rucklidge, W. J. (1993). 'Comparing images using the Hausdorff distance', *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15(9): 850–863.

Jacoby, N. and McDermott, J. H. (2017). 'Integer ratio priors on musical rhythm revealed cross-culturally by iterated reproduction', *Current Biology 6* 27(3): 359–370, https://doi.org/10.1016/j.cub.2016.12.031.

Kalish, M. L., Griffiths, T. L., and Lewandowsky, S. (2007). 'Iterated learning: Intergenerational knowledge transmission reveals inductive biases', *Psychonomic Bulletin & Review* 14: 288–294, https://doi.org/10.3758/BF03194066.

Kemp, C. and Regier, T. (2012). 'Kinship categories across languages reflect general communicative principles', *Science* 336: 1049–1054.

Kempe, V., Gauvrit, N., and Forsyth, D. (2015). 'Structure emerges faster during cultural transmission in children than in adults.' *Cognition* 136: 247–254.

Kempe, V., Gauvrit, N., Gibson, A., and Jamieson, M. (2019). 'Adults are more efficient in creating and transmitting novel signalling systems than children', *Journal of Language Evolution* 4(1): 44–70, https://doi.org/10.1093/jole/lzy012.

Kempe, V., Gauvrit, N., Panayotov, N., et al. (2021). 'Amount of learning and signal stability modulate emergence of structure and iconicity in novel signalling systems', *Cognitive Science* 45(11): [e13057], https://doi.org/10.1111/cogs.13057.

9780192886491_Book.indb 21 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:21

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Kirby, S. (2017). 'Culture and biology in the origins of linguistic structure', *Psychonomic Bulletin & Review* 24: 118–137, https://doi.org/10.3758/s13423-016-1166-7.

Kirby, S., Cornish, H., and Smith, K. (2008). 'Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language', *Proceedings of the National Academy of Sciences* 105(31): 10681–10686, https://doi.org/10.1073/pnas.0707835105.

Kirby, S., Tamariz, M., Cornish, H., and Smith, K. (2015). 'Compression and communication in the cultural evolution of linguistic structure', *Cognition* 141: 87–102, https://doi.org/10.1016/j.cognition.2015.03.016.

Knowlton, B. J. and Squire, L. R. (1994). 'The information acquired during artificial grammar learning', *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20: 79–91.

Levenshtein, V. I. (1966). 'Binary codes capable of correcting deletions, insertions, and reversals', *Soviet Physics-Doklady* 10(8): 707–710.

Lumaca, M. and Baggio, G. (2017). 'Cultural transmission and evolution of melodic structures in multi-generational signaling games', *Artificial Life* 23: 406–423, https://doi.org/10.1162/ARTL_a_00238.

Ma, W. and Fiveash, A. (2019). 'Spontaneous emergence of language-like and music-like vocalizations from an artificial protolanguage', *Semiotica* 229: 1–23, https://doi.org/10.1515/sem-2018-0139.

Mesoudi, A., Whiten, A., and Dunbar, R. (2006). 'A bias for social information in human cultural transmission', *British Journal of Psychology (London, England:* 1953) 97(3): 405–423, https://doi.org/10.1348/000712605X85871.

Miton, H., Wolf, T., Vesper, C., et al. (2020). 'Motor constraints influence cultural evolution of rhythm', *Proceedings of the Royal Society B* 287: 1937.

Motamedi, Y., Schouwstra, M., Smith, K., et al. (2019). 'Evolving artificial sign languages in the lab: From improvised gesture to systematic sign', *Cognition* 192: 103964.

Müller, T. F., Winters, J., Morisseau, T., et al. (2021). 'Colour terms: Native language semantic structure and artificial language structure formation in a large-scale online smartphone application', *Journal of Cognitive Psychology* 33(4): 357–378, https://doi.org/10.1080/20445911.2021.1900199.

Page, E. B. (1963). 'Ordered hypotheses for multiple treatments: A significance test for linear ranks', *Journal of the American Statistical Association* 58: 216–230.

Pelli, D. G., Burns, C. W., Farell, B., and Moore-Page, D. C. (2006). 'Feature detection and letter identification', *Vision Research* 46(28): 4646–4674, https://doi.org/10.1016/j.visres.2006.04.023.

Perfors, A. and Navarro, D. J. (2014). 'Language evolution can be shaped by the structure of the world', *Cognitive Science* 38: 775–793, https://doi.org/10.1111/cogs.12102.

Raviv, L., Meyer, A., and Lev-Ari, S. (2019). 'Compositional structure can emerge without generational transmission', *Cognition* 182: 151–164, https://doi.org/10.1016/j.cognition.2018.09.010.

Saldana, C., Kirby, S., Truswell, R., and Smith, K. (2019). 'Compositional hierarchical structure evolves through cultural transmission: An experimental study', *Journal of Language Evolution* lzz002, https://doi.org/10.1093/jole/lzz002.

Silvey, C., Kirby, S., and Smith, K. (2015). 'Word meanings evolve to selectively preserve distinctions on salient dimensions', *Cognitive Science* 39(1): 212–226, https://doi.org/10.1111/cogs.12150.

Smith, K. and Wonnacott, E. (2010). 'Eliminating unpredictable variation through iterated learning', *Cognition* 116(3): 444–449.

9780192886491_Book.indb 22 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:21

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Smith, K. (2018). 'The cognitive prerequisites for language: Insights from iterated learning', *Current Opinion in Behavioral Sciences* 21: 154–160, https://doi.org/10.1016/j.cobeha.2018.05.003.

Smith, K. and Dediu, D. (2020). 'Did partial hearing loss shape Australian phonology? An empirical test using spoken iterated learning', *The Evolution of Language: Proceedings of the 13th International Conference*. Nijmegen: Max Planck Institute for Psycholinguistics.

Tamariz, M. (2008). 'Exploring systematicity between phonological and context-cooccurrence representations of the mental lexicon', *The Mental Lexicon* 3(2): 259–278.

Tamariz, M. (2011). 'Linguistic structure evolves to match meaning structure', in L. Carlson, C. Hölscher, and T. Shipley (eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Tamariz, M. and Kirby, S. (2015). *Culture: Copying, Compression and Conventionality*. *Cognitive Science*, https://doi.org/10.1111/cogs.12144.

Tamariz, M. and Smith, A. D. M. (2008). 'Regularity in mappings between signals and meanings', in A. D. M. Smith, K. Smith, and R. Ferrer i Cancho (eds.), *The Evolution of Language (EVOLANG 7)*. Barcelona: World Scientific.

Tamariz, M., Brown, J. E., and Murray, K. M. (2010). 'The role of practice and literacy in the evolution of linguistic structure', in A. D. M. Smith, B. de Boer, M. Schouwstra, and K. Smith (eds.), *The Evolution of Language: Proceedings of the 8th International Conference on the Evolution of Language*. Utrecht: World Scientific Press.

Tamariz, M., Kirby, S., and Carr, J. W. (2016). 'Cultural evolution across domains: Language, technology and art', in A. Papafragou, D. Grodner, D. Mirman, and J. Trueswell (eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, 2765–2770. Philadelphia: Cognitive Science Society.

Tamariz, M., Roberts, S. G., Martínez, J. I., and Santiago, J. (2018). 'The interactive origin of iconicity', *Cognitive Science* 42(1): 334–349.

Theisen-White, C., Kirby, S., and Oberlander, J. (2011). 'Integrating the horizontal and vertical cultural transmission of novel communication systems', *Proceedings of the Annual Meeting of the Cognitive Science Society* 33(33), https://escholarship.org/uc/item/40t7578c

Thompson, B., Kirby, S., and Smith, K. (2016). 'Culture shapes the evolution of cognition', *PNAS* 113: 4530–4535, https://doi.org/10.1073/pnas.1523631113.

Tinits, P., Nölle, J., and Hartmann, S. (2017). 'Usage context influences the evolution of overspecification in iterated learning', *Journal of Language Evolution* 2(2): 148–159, https://doi.org/10.1093/jole/lzx011.

Verhoef, T. (2012). 'The origins of duality of patterning in artificial whistled languages', *Language and Cognition* 4(4): 357–380, https://doi.org/10.1515/langcog-2012-0019.

Verhoef, T. and Ravignani, A. (2021). 'Melodic universals emerge or are sustained through cultural evolution', *Frontiers in Psychology* 12: 668300, https://doi.org/10.3389/fpsyg.2021.668300.

Verhoef, T., Kirby, S., and de Boer, B. (2014). 'Emergence of combinatorial structure and economy through iterated learning with continuous acoustic signals', *Journal of Phonetics* 43: 56–68, https://doi.org/10.1016/j.wocn.2014.02.005.

Verhoef, T., Walker, E., and Marghetis, T. (2016). 'Cognitive biases and social coordination in the emergence of temporal language', *The 38th Annual Meeting of the Cognitive Science Society* 2615–2620.

Wasielewski, H. (2014). 'Imitation is necessary for cumulative cultural evolution in an unfamiliar, opaque task', *Human Nature* 25: 161–179, https://doi.org/10.1007/s12110-014-9192-5.

9780192886491_Book.indb 23 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:22

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Winters, J., Kirby, S., and Smith, K. (2015). 'Languages adapt to their contextual niche', *Language and Cognition* 7(3): 415–449, https://doi.org/10.1017/langcog.2014.35.

Xu, J., Griffiths, T. L., and Dowman, M. (2010). 'Replicating color term universals through human iterated learning', in *Proceedings of the Annual Meeting of the Cognitive Science Society* 32(32), https://escholarship.org/uc/item/64t7z802.

9780192886491_Book.indb 24 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:22

CHAPTER 2

ARTIFICIAL LANGUAGE LEARNING

SHIRA TAL, INBAL ARNON, AND JENNIFER CULBERTSON

2.1 Introduction

All natural languages share a basic set of functions: they have to be learned by children (and sometimes adults); to serve communicative goals; and to be used by a community (Hockett 1960; Bates and MacWhinney 1989; Christiansen and Kirby 2003; Hawkins 2004; Christiansen and Chater 2016; Gibson et al. 2019). Importantly, all natural languages are learned and used by the same human mind and body: the interplay between the functions of language, and the cognitive and perceptual biases of the human learning and using it, could potentially shape language structure over time. Indeed, cognitive and perceptual biases offer empirically testable explanations for recurrent typological patterns: linguistic patterns that are more frequent across the world's languages may be more learnable, easier to perceive, process, and produce, or more efficient for communication. Importantly, neither typological nor theoretical linguistics research alone is sufficient to test these explanations. First, typological data are often sparse and confounded by genetic and areal relationships between languages, rendering only limited support for any given account. More importantly, typological data cannot address the *cognitive* mechanisms postulated to be involved in processes of language evolution: corpora and grammars of languages may show the outcomes of cognitive processes but they cannot provide causal evidence for them. In recent years, new experimental and computational paradigms have been used to investigate the hypothesized link between typological patterns and cognitive or perceptual biases (e.g. Tamariz and Kirby 2016; Smith 2022). For instance, by showing that a common linguistic pattern is also easier to learn than others, we can provide direct behavioural evidence linking cognitive biases to language typology. In this chapter, we focus on artificial language

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

learning experiments, one of the most common methods used to investigate how the human mind impacts language structure. Artificial language learning experiments allow us to directly test biases active during learning by asking how well participants learn, generalize, or make inferences about different linguistic structures, including the ways in which they change the input language they were exposed to. This method allows researchers to test the impact of learning biases (or constraints) on language structure in ways that are not possible when looking at natural language acquisition outside the lab.

Artificial language learning (ALL) is an umbrella term used to describe a host of experimental paradigms. While these paradigms differ from one another (see section 2.2), they all share a core design feature: participants are exposed to miniature novel linguistic systems, comprised of strings, words, or phrases.[1] These miniature languages allow experimenters to control the amount and type of input that participants are exposed to, creating a kind of petri dish where factors of interest can be isolated and studied. For example, researchers can compare the learnability of different linguistic features, such as different word orders (e.g. learning an SVO language vs. a VSO language), holding the rest of the system constant. This type of comparison is virtually impossible to do with natural languages in the wild. First, natural languages differ from one another in a multitude of ways, making it very difficult to determine the effect of a single feature on learnability. Moreover, it is difficult to compare different *learners* of natural languages; even learners of the same language will differ in the amount and kind of exposure they have to that language. Thus, by exposing all participants to a novel language in the same way, and under the same conditions, researchers can minimize confounds and test hypotheses about the link between biases and language structure.

ALL experiments have been a key experimental tool in studying language acquisition for many years. In a seminal study, Saffran et al. (1996) showed that infants were able to use distributional cues to segment continuous linguistic speech following two minutes of exposure to an artificial language. This study (and numerous studies since, see Saffran 2020; Isbilen and Christiansen 2022) showed that infants are capable of extracting language-relevant information using statistical learning—by tracking and extracting information about the co-occurrence of various elements. These findings have far-reaching implications for language acquisition, but also for language evolution. In particular, they suggest that important aspects of language can be learned by domain-general mechanisms, and not necessarily language-specific ones. Along the same lines, artificial language learning studies demonstrating the advantage of multiword chunks (for a review, see Arnon 2021) shed light not only on mechanisms of language acquisition (Arnon and Christiansen 2017; Skarabela et al. 2021) but also on the selection of language structures over time (Isbilen and Christiansen 2020): the learning and memory advantage of reusable chunks is argued to lead to repeated chunks in language structure.

[1] Such experiments are also sometimes referred to as Artificial Grammar Learning, reflecting the fact that the systems participants learn do not necessarily have to be meaningful, or even linguistic, but typically involve some set of rules or a grammar to be learned (e.g. Reber 1967, 1989).

9780192886491_Book.indb 26 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:22

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

While ALL is widely used in language acquisition research, and many ALL studies in this field have important implications for language evolution, in this chapter, we focus on ALL experiments that test the link between cognitive or perceptual biases and recurrent typological patterns (often called 'language universals'). In recent decades, a number of specific cognitive mechanisms have been investigated using ALL. For example, many prevalent cross-linguistic patterns across different levels of linguistic analysis are argued to be affected by a bias for simplicity—speakers' preference for systems that can be described more concisely (in information-theoretic terms: with more compressible descriptions; Culbertson and Kirby 2016). There is a growing body of evidence from ALL studies, showing that learners have a bias for simpler systems (at the phonological level, Moreton and Pater 2012; at the morphological level, Mansfield et al. 2022; at the syntactic level, Culbertson et al. 2012, 2020, Culbertson and Newport 2015). Other learning biases and pressures potentially active during learning that have been investigated using ALL include regularity (Hudson Kam and Newport 2005, 2009; Perfors 2012; Samara et al. 2017; Schwab et al. 2018; Ferdinand et al. 2019; Saldana, Smith, et al. 2021), iconicity (Brand et al. 2018), efficiency (Fedzechkina et al. 2012, 2017; Kanwal et al. 2017; Kurumada and Grimm 2019; Fedzechkina and Jaeger 2020), predictability (Kurumada et al. 2013; Seyfarth et al. 2014; Hendrickson and Perfors 2019; Johnson et al. 2020; Lavi-Rotbain and Arnon 2022), phonological naturalness (White 2014; Mintz et al. 2018; Martin and Peperkamp 2020; Martin and White 2021), perceptual similarity (Hupp et al. 2009; Bruening et al. 2012; Martin and Culbertson 2020), (morpho-)syntactic naturalness (e.g. structural iconicity, Culbertson and Adger 2014; Martin et al. 2020; Saldana, Oseki, et al. 2021), as well as biases in conceptual structure (e.g. connectedness, Chemla, Buccola, et al. 2019; Chemla, Dautriche, et al. 2019), category learning (Cristia and Seidl 2008; Moreton et al. 2015; Culbertson et al. 2019; Carr et al. 2020), and semantic representations (Maldonado et al. 2022; Maldonado and Culbertson 2020; Saratsli et al. 2020).

In what follows, we will provide a methodological guideline for using ALL experiments to investigate the link between language universals and learning. We begin with presenting three main paradigms of ALL experiments that are used in language evolution research. We will briefly discuss how they relate to other experimental methods used in this field. We will then describe some key practical considerations of running ALL studies. We will end by discussing the limitations of this method, and outlining important areas that should be developed in future research to push this method forward.

2.2 Key paradigms

There are three main ALL paradigms that are used to study the link between learning biases and language structure: Ease of Learning; Extrapolation; and Regularization

9780192886491_Book.indb 27 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:22

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

(Culbertson 2012, 2023). Below, we describe each paradigm along with two examples of studies using it in different linguistic domains.

In the Ease of Learning paradigm, participants are trained on artificial language systems that differ in some features of interest. Following exposure, participants are tested on what they have learned. The outcome measure is how well participants learn the language *as it appeared in their input*. Crucially, learning accuracy is compared across the different systems learned. As with other ALL methods, studies using Ease of Learning can differ in a number of specific design dimensions, including whether novel items are used in testing (i.e. testing generalization), what is tested (e.g. comprehension or production), and how it is tested (force-choice, typing, oral production, etc.). Tal and Arnon (2022) use this paradigm to investigate whether a redundant morpheme (i.e. a morpheme that adds no information) would help child learners to acquire thematic role assignment (i.e. who-did-what-to-whom) in an artificial language. They compared learning of two languages, both with fixed word order, but one which also contained a redundant object case marker. Children were tested on their comprehension of novel sentences in the language using a picture-matching task. They found that children exposed to the language with the redundant case marker were better at learning thematic role assignment. Tal and Arnon (2022) argue that the benefit of redundant morphemes during learning may help explain the cross-linguistic prevalence of redundancy, despite the added complexity (Leufkens 2020). Maldonado and Culbertson (2020) used this paradigm to investigate the learnability of different person systems. They compared learning of systems with different homophony patterns: ones that have a feature-based homophony pattern (e.g. having identical forms for entities that share the same number) and ones that have a random homophony pattern (having identical forms for entities that do not share any feature). They found that participants who were exposed to a feature-based homophony pattern—even one that is not attested in their own language—learned the language better than participants exposed to a random homophony pattern. These findings show that rules that can be characterized on the basis of features are easier to learn than rules that group together a set of distinct segments (see Saldana et al. 2022 for a similar account in morphology).

In the Extrapolation paradigm (sometimes called the Poverty of the Stimulus paradigm; Culbertson 2023), participants are trained on input that is compatible with two (or more) hypotheses of interest. Testing involves targeted extrapolation beyond the input, designed to reveal which of the competing hypotheses participants have inferred. For example, Martin and colleagues (2020) used this method to investigate cognitive biases in the relative ordering of modifiers in the noun phrase (e.g. *these three red roses*). Participants were trained on phrases in a novel language that were composed of a noun and a single post-nominal modifier, either a demonstrative (akin to 'feather that') or an adjective (akin to 'feather red'). In testing, they were asked to describe images which required using both modifiers together. They were interested in whether participants would produce adjectives closer to the noun than demonstratives (i.e. something like 'feather red that')—a cross-linguistically common pattern (Greenberg 1963;

9780192886491_Book.indb 28 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:22

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Dryer 2018)—or vice versa (i.e. something like 'feather that red'). Participants indeed preferred the cross-linguistically common order, suggesting that the typological distribution is driven by a cognitive bias (for corroborating experimental evidence and discussion of the nature of this bias, see Culbertson and Adger 2014; Culbertson et al. 2020). Martin and White (2021) used this method to investigate whether learners have a bias for vowel harmony (a phonological alternation where vowels within a word assimilate on some feature like backness or roundness). They trained participants on a language with vowel harmony or vowel disharmony between stems and a single suffix, and then tested them on unseen combinations of multiple suffixes in order to see whether they extrapolated harmony or disharmony. They found that learners readily extrapolated the vowel harmony pattern but not disharmony. This supports the hypothesis that prevalence of vowel harmony in the world languages (Rose and Walker 2011) is driven by a learning advantage.

In the Regularization paradigm, participants are exposed to a language system that contains variability, e.g. more than one alternative used to describe the same meaning. Of interest is whether and how participants' output differs from their input. In some cases, researchers are interested in the conditions under which participants will overuse one of the alternatives in the input compared to its input rate—called regularization (Hudson Kam and Newport 2009; Perfors 2012; Ferdinand et al. 2019). In their classic paper, Hudson Kam and Newport (2009) teach adults and children languages in which nouns occur variably with different determiners. They show that children are more likely to regularize the variation than adults, and argue that this highlights the importance of children in driving regularity in language. In other cases, researchers are interested in whether participants' output reflects a preference for one particular alternative (compared to other alternatives). Culbertson and colleagues (2020) used the Regularization paradigm to investigate a well-known typological universal: word order harmony. This is the tendency for consistent ordering of heads and dependents, either within or across phrases. For example, nouns tend to appear in a consistent position relative to their modifiers (e.g. 'these/three/empty bottles'; Greenberg 1963). Culbertson and colleagues (2020) exposed participants to noun phrases comprised of either nouns and numerals or nouns and adjectives. The order of nouns and modifiers was variable in the languages participants learned, however, some participants were randomly assigned to a harmonic condition—where the dominant word order had adjectives and numerals in the same position relative to the noun—or a non-harmonic condition—where the dominant word order had adjectives and numerals appearing on different sides of the noun. At test, participants had to produce noun phrases in the novel language. They tested native English-speaking adults and children, but also native French- and Hebrew-speaking adults and children, whose language uses a non-harmonic order of modifiers with numerals before the noun but adjectives after. All participants, even those whose native language was non-harmonic (including monolingual children), showed a preference for harmonic word orders (Culbertson, Franck, et al. 2020; for further ALL experiments supporting this account, see Culbertson et al. 2012, Culbertson and Newport 2015, Culbertson et al. 2024, Wang et al., in press). These findings suggest

9780192886491_Book.indb 29 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:22

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

that the origin of this typological universal lies in a cognitive bias active during learning. In particular, a bias for simplicity predicts that a language with a single general ordering rule is easier to learn than one with several different rules (Culbertson and Kirby 2016; see section 2.1).

The three artificial language learning paradigms described in this section are the most popular basic designs used to link cognitive biases to typology. The examples we have highlighted here involve a setup in which participants learn part of a language system, and are then tested on their ability to reproduce that language, or use what they have learned in new ways. However, it is worth noting that all these designs can, in principle, be combined with more complex methods like iterated learning or communication experiments (as discussed in this volume, Chapters 1 and 3), and the input can vary across domains and modalities (as discussed in this volume, Chapter 9, see also Culbertson et al. 2024).

2.3 Designing an ALL experiment

ALL is an umbrella term for different paradigms, and while there are several guiding principles that are important to consider, there is by no means one recipe to create an artificial language. Different experiments, even ones that employ the same paradigm, use languages that vary greatly in their complexity, the size of the lexicon, and their artificiality (that is, how different they are from participants' native language). These design decisions depend on the research question, the population tested, as well as other practical considerations. In what follows, we will outline important issues to consider when designing an artificial language.

The semantics of the language. Some ALL experiments teach participants novel words and phrases which are used to refer to known referents or actions. For example, in Tal and Arnon's (2022) study on redundant morphology, phrases in the language denote common actions involving recognizable characters, such as a painter kicking a chef. Other studies make use of *novel* actions or referents—ones that participants are unfamiliar with. For example, the nouns taught in Culbertson et al. (2020) refer to novel objects (see Figure 2.1). Since participants have no readily available label for novel objects in their native language, this method is useful for discouraging participants from translating phrases directly from their own language. Finally, in some ALL experiments, participants learn languages with no semantics at all. For example, Culbertson et al. (2024) teach participants sequences of stimuli made up of shapes, non-linguistic sounds, and sequences of tactile pulses. This kind of design is also standard in statistical learning experiments, where researchers are often interested in how learners segment speech into words, irrespective of any extralinguistic cues. In these experiments, participants therefore listen to streams of novel sounds, but these sounds typically do not correspond to any referents (see Erickson and Thiessen 2015 for a review).

9780192886491_Book.indb 30 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:22

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

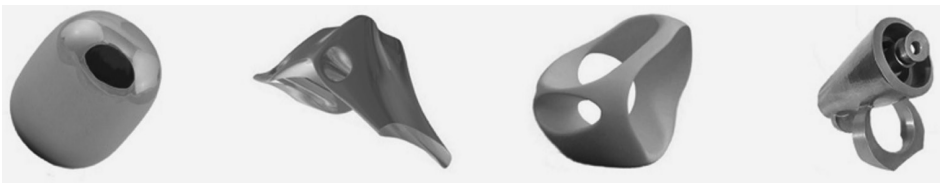


FIGURE 2.1 Stimuli used in Culbertson et al. (2020) to denote novel objects that participants are unfamiliar with.

Source: Culbertson et al. (2020).

How artificial should the language be? In principle, the language should be as artificial—that is, as unfamiliar to participants—as possible. The farther the novel language is from participants' native language, the more confident researchers can be that participants' behaviour is driven by biases that are independent of their specific linguistic experience (see further discussion on that below). To this end, the lexicon and other features of the language should ideally be as different as possible from participants' native language, while still being learnable (e.g. not containing sounds that participants cannot perceive or produce). For example, in Martin et al.'s (2020) study on modifier order in the noun phrase, 'eyey puku' could mean 'red feather' and 'eyey taka' could mean 'that feather'. This study was run with English speakers, but these words are fairly dissimilar to English and crucially the word order is the opposite of English word order.[2] In some cases, however, there might be reasons to opt for a non-artificial lexicon, or to use other linguistic features that are less unfamiliar to participants. For example, if the research question does not target word-learning but some other grammatical feature (e.g. word order), researchers might want to ease the pressure of learning the lexicon and limit the challenging parts of learning to the higher-level patterns of interest. A prominent example of such case is when running ALL experiments on children. As in many cognitive experiments, ALL experiments are typically more difficult for children than for adults (Raviv and Arnon 2018; Culbertson et al. 2019; Tal and Arnon 2022), and children's patience to participate is generally limited. For these reasons, ALL experiments with children have to be adapted to their needs, by being short and not too taxing (for a review on ALL experiments with children, see Culbertson and Schuler 2019). However, even with adults, a novel lexicon can be difficult to learn, and may be unnecessary. One straightforward way to simplify the lexicon is to use words that are very similar to participants' native language. For example, in Tal and Arnon's (2022) study, done on Hebrew-speaking children (and adults), the lexicon was composed of Hebrew nouns and novel suffixes (e.g. 'zayarig', the English equivalent of which is 'painterig', to describe a painter; see Merkx et al.

[2] Rather than inventing an artificial lexicon, another option is to use a lexicon of a natural language that participants are not familiar with. For example, Brooks et al. (2011) exposed English-speaking participants to Russian nouns in order to investigate whether similar word endings (e.g. *pechka* and *myshka*) can benefit learning of gender classes.

9780192886491_Book.indb 31 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:22

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

2011 for a similar example combining English stems with novel affixes). Another way to achieve the same goal is to use iconic lexical items for referents. For example, in Fehér et al. (2019), the nouns refer to animals and have iconic names, such as 'oinko' for pig. Finally, some experiments even use a completely familiar lexicon (i.e. words in the participants' native language) with an unfamiliar syntax (e.g. different word order, as in Culbertson and Adger 2014; Meisezahl et al., in press).[3]

How complicated should the language be? ALL experiments are a simplified and abstracted version of learning a language. While this simplification is also an advantage, allowing researchers to eliminate confounding factors, it is always possible that these experiments can abstract away factors that are important and relevant to the phenomenon at hand. The degree of language complexity should, therefore, be carefully tailored to the case at hand. Indeed, the complexity of the language that participants are trained on varies widely between studies. Aside from the novelty of the lexicon, other variable factors are the size of the lexicon and the complexity of the grammar. The complexity of the language will have a knock-on impact on the amount of training participants are likely to need. As mentioned above, child participants require shorter and easier tasks compared to adult participants. For this reason, even when tested on the same phenomena, children are usually tested on a smaller-scale version of the language that adults are tested on. For example, in Culbertson, Franck, et al.'s (2020) study on modifier order in the noun phrase, children (6–8 years old) and adults went through the same experimental paradigm. However, adults learned a language that consisted of ten nouns and ten modifiers, whereas children learned a language with only four nouns and six modifiers. If the exposure is done over several days (Hudson Kam and Newport 2009; Fedzechkina et al. 2012), the lexicon can be larger compared to experiments that last only ten to fifteen minutes (Culbertson and Adger 2014; Tal and Arnon 2020, 2022). The length of the experiment may also depend on financial constraints or experimental settings (e.g. are participants conducting the study online, in a dedicated lab, or as part of a different activity, such as a visit to a science museum?).

In general, researchers tend to limit the complexity of the language as much as possible in light of the questions of interest. For example, if the main research question is under what conditions people regularize linguistic variation, then there may be no need for the language to have complex sentences—phrases with a word and a plural marker will do (e.g. Smith and Wonnacott 2010). If the main research question is how people prefer to order nouns, adjectives, and demonstratives, then having plural marking on nouns would just make the language more difficult to learn (e.g. Martin et al. 2020). But regularization experiments might require a critical mass of different words occurring

[3] There are also a number of experiments using artificial language systems that are purposefully non-language-like (e.g. Little et al. 2017), or utilize alternative modalities for communication (e.g. Motamedi et al. 2022). See Naegeli and Schouwstra (Chapter 9, this volume) on Silent Gesture, and Tamariz and Papa (Chapter 1, this volume) on Iterated Learning for additional examples.

9780192886491_Book.indb 32 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:22

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

(or not) with a marker, whereas an experiment on word order preferences might only need to have two words of each category.

How to minimize the influence of participants' natural language? One of the core limitations of ALL experiments is that they are done on participants who already speak one or more languages. Participants are affected by their native language when learning an artificial language (Siegelman et al. 2018; Tang and Baer-Henney 2023), just as they are during natural second language learning. This is a potentially major confound in ALL experiments: it can be difficult to dissociate participants' behaviour in the task and their prior language experience. There are several possible ways to deal with this issue. First, researchers can choose participants whose language does not feature the pattern of interest. For example, Fedzechkina et al. (2012) investigate the emergence of differential case marking in English speakers, whose language (by and large) does not have case marking. However, this is not always possible. For example, many artificial language learning experiments in the field of language evolution have investigated word order, and all languages (more or less) have particular ways of ordering words. In this case, the best technique is to compare populations whose native language features different orders; ideally including a population whose language goes against the bias of interest. For example, Culbertson et al. (2020) investigate whether participants have a bias for harmonic word orders. Crucially, this study was done on speakers of languages with non-harmonic word orders—Hebrew and French. Moreover, this experiment was done on monolingual Hebrew- and French-speaking children, who had no knowledge of English (which has a harmonic word order).

Notably, most of the ALL studies to date have been done on English-speaking participants. This problem is, of course, not unique to ALL experiments but poses a serious limitation to psychology and cognitive science more broadly (Henrich et al. 2010; Nielsen et al. 2017). In general, testing mainly English-speaking participants on university campuses or on crowdsourcing platforms like Prolific leads to a very homogenous sample that is not necessarily generalizable to other communities. However, it is arguably even more problematic for ALL experiments: if most of our data come from speakers of just one language, that limits considerably any conclusions that could be made about the relation between *general* cognitive biases and *general* principles of language evolution and change. Testing speakers of other languages is a crucial component in the aim to dissociate participants' specific experience with their language and universal cognitive biases that have the potential to influence language typology.

Finally, beyond the possible influence of participants' *specific* language experience, participants' learning can also be influenced by having a language—any language. Participants' extensive experience with language systems, along with their rudimentary basic properties (e.g. compositionality, duality of patterning, etc.) could challenge researchers' ability to investigate how these properties emerge. In the next section, we outline exciting new directions in ALL that can also address this limitation.

9780192886491_Book.indb 33 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:22

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

2.4 Next directions in ALL research

ALL experiments are a powerful tool to study the relation between cognitive biases and language change. Though ALL has been used for a long time in the cognitive sciences, it is still relatively new in the field of language evolution. We would like to highlight several exciting new directions in investigating the link between cognitive biases and language structure. These directions are starting to appear in recent work, and we believe future work should push them forward even more.

First and foremost, as already hinted at earlier, most of the ALL studies to date have been done on English-speaking participants. This tendency is, however, starting to change, revealing a richer understanding of the cognitive biases that underlie linguistic structures on the one hand, and the impact of language experience on learning on the other. Existing work comparing learning across languages demonstrates that while in some cases findings coming from English-speaking participants are replicated for other languages, in other cases they are not. For example, as mentioned above, participants' bias for learning harmonic patterns is attested not only for speakers of a harmonic language like English (Culbertson et al. 2012) but also for speakers of non-harmonic languages like French and Hebrew (Culbertson, Franck, et al. 2020). Importantly, this is not always the case. The well-known *suffixing preference*—the cross-linguistic preference to form complex words using suffixes rather than prefixes—provides a case in point (Greenberg 1957; Dryer 2013). This tendency has often been attributed to a general bias to perceive sequences that differ at their ends (similar to suffixed words) as more similar than sequences that differ at their beginnings (e.g. Hupp et al. 2009). However, until recently, the evidence for this claim came exclusively from speakers of English—a language that uses extensive suffixing. When the same experiment was run on speakers of Kîîtharaka—a Bantu language that relies heavily on prefixing—these findings did not replicate (Martin and Culbertson 2020), calling into question whether this bias can indeed underlie the suffixing preference.

Another direction for future research is to investigate the link between cognitive biases and language typology and change beyond adult learners. Very few ALL experiments have investigated whether and how cognitive biases that are postulated to impact language change might differ in adults and children. This is especially problematic since the small number of experiments that have been done with children showed that children and adults show different learning biases (Hudson Kam and Newport 2009; Samara et al. 2017; Raviv and Arnon 2018; Schwab et al. 2018; Culbertson et al. 2019; Tal and Arnon 2020). Differences between children and adults have important implications for language change. First, these differences can inform the longstanding debate about who is driving language change. While some theories argue that child learners drive language change over time by acquiring grammars that are (slightly) different from the grammars of the previous generation (e.g. Hudson Kam and Newport 2009; Cournane 2019), others claim that languages are instead shaped primarily or

9780192886491_Book.indb 34 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:22

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

exclusively by adolescents and adults (e.g. Labov 2007). Second, comparing the cognitive biases of children and adults allows for a better understanding of the nature of these biases: while some biases might be already attested early on in development, others might rely on greater experience with the world or certain aspects of it. Indeed, one way to separate existing linguistic or world knowledge from experience-independent biases is to test not only young children, but also infants (Saffran and Thiessen 2003; Cristià and Seidl 2008; Gómez et al. 2014; White and Sundara 2014).

Another way to tease apart language experience from general cognitive biases, and an important path to take in order to get a fuller evolutionary picture, is to conduct cross-species comparisons using ALL experiments. Some recent cross-species studies provide exciting evidence, both for the plausibility of such an endeavour and for the insights that it can provide. For example, words across the world's languages tend to express connected meanings: if a word can be used to convey two meanings, it will also apply to meanings that fall 'between' those two in conceptual space. For instance, nouns tend to refer to 'connected' classes such as different types of 'birds' but not to 'unconnected' classes such as 'either table or shoe'. This tendency is argued to derive from a learning bias favouring connected words (Chemla, Buccola, et al. 2019). Remarkably, a recent study has used an adapted ALL task to show that this bias is also attested in baboons, supporting the idea that this is a cognition-general bias, not specific to language, and shared with other species (Chemla, Dautriche, et al. 2019). Conducting cross-species studies is an important and exciting avenue to push the field forward.

To summarize, ALL experiments are an important tool for researchers aiming to test hypothesized links between cognition and language structure and typology. Studies to date have targeted a range of phenomena, from vowel harmony, to differential case marking, to personal pronoun systems. To strengthen the evidence coming from this paradigm even further, future work should broaden its scope to include populations coming from different cultures and speaking different languages, and to compare across development, and across species. While this method has several limitations that need to be kept in mind when designing studies, the evidence coming from ALL experiments is a powerful source of converging evidence, alongside typological samples, other experimental methods (e.g. silent gesture, iterated learning; see Chapters 9 and 1, this volume), and computational modelling.

References

Arnon, I. (2021). 'The Starting Big approach to language learning', *Journal of Child Language* 48(5): 937–958, https://doi.org/10.1017/S0305000921000386.

Arnon, I. and Christiansen, M. H. (2017). 'The role of multiword building blocks in explaining L1–L2 differences', *Topics in Cognitive Science* 9(3): 621–636, https://doi.org/10.1111/tops.12271.

Bates, E. and MacWhinney, B. (1989). 'Functionalism and the competition model', in B. Macwhinney and E. Bates (eds.), *The Crosslinguistic Study of Sentence Processing*, 3–76. New York: Cambridge University Press.

9780192886491_Book.indb 35 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:22

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Brand, J., Monaghan, P., and Walker, P. (2018). 'The changing role of sound-symbolism for small versus large vocabularies', *Cognitive Science* 42: 578–590, https://doi.org/10.1111/cogs.12565.

Brooks, P. J., Kempe, V., and Donachie, A. (2011). 'Second language learning benefits from similarity in word endings: Evidence from Russian', *Language Learning* 61(4): 1142–1172, https://doi.org/10.1111/j.1467-9922.2011.00665.x.

Bruening, P. R., Brooks, P. J., Alfieri, L., et al. (2012). 'Children's tolerance of word-form variation', *Child Development Research* 2012(1): 401680 https://doi.org/10.1155/2012/401680.

Carr, J. W., Smith, K., Culbertson, J., and Kirby, S. (2020). 'Simplicity and informativeness in semantic category systems', *Cognition* 202(July): 104289, https://doi.org/10.1016/j.cognition.2020.104289.

Chemla, E., Buccola, B., and Dautriche, I. (2019). 'Connecting content and logical words', *Journal of Semantics* 36(3): 531–547, https://doi.org/10.1093/jos/ffz001.

Chemla, E., Dautriche, I., Buccola, B., and Fagot, J. (2019). 'Constraints on the lexicons of human languages have cognitive roots present in baboons (Papio papio)', *Proceedings of the National Academy of Sciences of the United States of America* 116(30): 14926–14930, https://doi.org/10.1073/pnas.1907023116.

Christiansen, M. H. and Chater, N. (2016). *Creating Language: Integrating Evolution, Acquisition, and Processing*. Cambridge, MA: MIT Press.

Christiansen, M. H. and Kirby, S. (2003). 'Language evolution: Consensus and controversies', *Trends in Cognitive Sciences* 7(7): 300–307, https://doi.org/10.1016/S1364-6613(03)00136-0.

Cournane, A. (2019). 'A developmental view on incrementation in language change', *Theoretical Linguistics* 45(3–4): 127–150, https://doi.org/10.1515/tl-2019-0010.

Cristià, A. and Seidl, A. (2008). 'Is infants' learning of sound patterns constrained by phonological features?', *Language Learning and Development* 4(3): 203–227, https://doi.org/10.1080/15475440802143109.

Culbertson, J. (2012). 'Typological universals as reflections of biased learning: Evidence from artificial language learning', *Linguistics and Language Compass* 6: 310–329, https://doi.org/10.1002/lnc3.338.

Culbertson, J. (2023). 'Artificial language learning', in J. Sprouse (ed.), *Oxford Handbook of Experimental Syntax*, 271–300. Oxford: Oxford University Press.

Culbertson, J. and Adger, D. (2014). 'Language learners privilege structured meaning over surface frequency', *Proceedings of the National Academy of Sciences of the United States of America* 111(16): 5842–5847, https://doi.org/10.1073/pnas.1320525111.

Culbertson, J. and Kirby, S. (2016). 'Simplicity and specificity in language: Domain-general biases have domain-specific effects', *Frontiers in Psychology* 6: 1–11, https://doi.org/10.3389/fpsyg.2015.01964.

Culbertson, J. and Newport, E. L. (2015). 'Harmonic biases in child learners: In support of language universals', *Cognition* 139: 71–82, https://doi.org/10.1016/j.cognition.2015.02.007.

Culbertson, J. and Schuler, K. (2019). 'Artificial language learning in children', *Annual Review of Linguistics* 5: 353–373.

Culbertson, J., Smolensky, P., and Legendre, G. (2012). 'Learning biases predict a word order universal', *Cognition* 122(3): 306–329, https://doi.org/10.1016/j.cognition.2011.10.017.

Culbertson, J., Jarvinen, H., Haggarty, F., and Smith, K. (2019). 'Children's sensitivity to phonological and semantic cues during noun class learning: Evidence for a phonological bias', *Language* 95(2): 268–293, https://doi.org/10.1353/lan.2019.0031.

9780192886491_Book.indb 36 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:22

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Culbertson, J., Franck, J., Braquet, G., et al. (2020). 'A learning bias for word order harmony: Evidence from speakers of non-harmonic languages', *Cognition* 204(June): 104392, https://doi.org/10.1016/j.cognition.2020.104392.

Culbertson, J., Schouwstra, M., and Kirby, S. (2020). 'From the world to word order: Deriving biases in noun phrase order from statistical properties of the world', *Language* 96(3): 696–717.

Culbertson, J., Compostella, A., and Kirby, S. (2024). 'Language structure reflects biases in pattern learning across domains and modalities', *Quarterly Journal of Experimental Psychology*: 17470218241282404.

Dryer, M. S. (2013). 'Prefixing vs. suffixing in inflectional morphology', in M. S. Dryer and M. Haspelmath (eds.), *The World Atlas of Language Structures Online*. Leipzig, Germany: Max Planck Institute for Evolutionary Anthropology.

Dryer, M. S. (2018). 'On the order of demonstrative, numeral, adjective, and noun', *Language* 94(4): 798–833, https://doi.org/10.1353/lan.2018.0054.

Erickson, L. C. and Thiessen, E. D. (2015). 'Statistical learning of language: Theory, validity, and predictions of a statistical learning account of language acquisition', *Developmental Review* 37: 66–108, https://doi.org/10.1016/j.dr.2015.05.002.

Fedzechkina, M. and Jaeger, T. F. (2020). 'Production efficiency can cause grammatical change: Learners deviate from the input to better balance efficiency against robust message transmission', *Cognition* 196: 104115.

Fedzechkina, M., Jaeger, T. F., and Newport, E. L. (2012). 'Language learners restructure their input to facilitate efficient communication', *Proceedings of the National Academy of Sciences* 109(44): 17897–17902, https://doi.org/10.1073/pnas.1215776109.

Fedzechkina, M., Newport, E. L., and Jaeger, T. F. (2017). 'Balancing effort and information transmission during language acquisition: Evidence from word order and case marking', *Cognitive Science* 41(2): 416–446, https://doi.org/10.1111/cogs.12346.

Fehér, O., Ritt, N., and Smith, K. (2019). 'Asymmetric accommodation during interaction leads to the regularisation of linguistic variants', *Journal of Memory and Language* 109: 104036.

Ferdinand, V., Kirby, S., and Smith, K. (2019). 'The cognitive roots of regularization in language', *Cognition* 184: 53–68, https://doi.org/10.1016/j.cognition.2018.12.002.

Gibson, E., Futrell, R., Piandadosi, S. T., et al. (2019). 'How efficiency shapes human language', *Trends in Cognitive Sciences* 23(5): 389–407, https://doi.org/10.1016/j.tics.2019.02.003.

Gómez, D. M., Berent, I., Benavides-Varela, S., et al. (2014). 'Language universals at birth', *Proceedings of the National Academy of Sciences* 111(16): 5837–5841, https://doi.org/10.1073/pnas.1318261111.

Greenberg, J. (1957). 'Order of affixing: A study in general linguistics', in J. Greenberg (ed.), *Essays in Linguistics*, 86–94. Chicago, IL: University of Chicago Press.

Greenberg, J. (1963). 'Some universals of grammar with particular reference to the order of meaningful elements', in J. Greenberg (ed.), *Universals of Language*, 73–113. Cambridge, MA: MIT Press.

Hawkins, J. A. (2004). *Efficiency and Complexity in Grammars*. Oxford: Oxford University Press.

Hendrickson, A. T. and Perfors, A. (2019). 'Cross-situational learning in a Zipfian environment', *Cognition* 189(February): 11–22, https://doi.org/10.1016/j.cognition.2019.03.005.

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). 'The weirdest people in the world?', *Behavioral and Brain Sciences* 33: 61–83, https://doi.org/10.1017/S0140525X0999152X.

Hockett, C. F. (1960). 'The origin of speech man', *Scientific American* 203(3): 88–97.

9780192886491_Book.indb 37 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:22

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Hudson Kam, C. L. and Newport, E. (2005). 'Regularizing unpredictable variation: The roles of adult and child learners in language formation and change', *Language Learning and Development* 1(2): 151–195, https://doi.org/10.1207/s15473341lld0102_3.

Hudson Kam, C. L. and Newport, E. L. (2009). 'Getting it right by getting it wrong: When learners change languages', *Cognitive Psychology* 59(1): 30–66, https://doi.org/10.1016/j.cogpsych.2009.01.001.

Hupp, J. M., Sloutsky, V. M., and Culicover, P. W. (2009). 'Evidence for a domain-general mechanism underlying the suffixation preference in language', *Language and Cognitive Processes* 24(6): 876–909, https://doi.org/10.1080/01690960902719267.

Isbilen, E. S. and Christiansen, M. H. (2020). 'Chunk-based memory constraints on the cultural evolution of language', *Topics in Cognitive Science* 12(2): 713–726, https://doi.org/10.1111/tops.12376.

Isbilen, E. S. and Christiansen, M. H. (2022). 'Statistical learning of language: A meta-analysis into 25 years of research', *Cognitive Science* 46(9): e13198, https://doi.org/10.1111/cogs.13198.

Johnson, T., Culbertson, J., Rabagliati, H., and Smith, K. (2020). *Assessing Integrative Complexity as a Predictor of Morphological Learning Using Neural Networks and Artificial Language Learning*, https://psyarxiv.com/yngw9/.

Kanwal, J., Smith, K., Culbertson, J., and Kirby, S. (2017). 'Zipf's law of abbreviation and the principle of least effort: Language users optimise a miniature lexicon for efficient communication', *Cognition* 165: 45–52.

Kurumada, C. and Grimm, S. (2019). 'Predictability of meaning in grammatical encoding: Optional plural marking', *Cognition* 191: 103953, https://doi.org/10.1016/j.cognition.2019.04.022.

Kurumada, C., Meylan, S. C., and Frank, M. C. (2013). 'Zipfian frequency distributions facilitate word segmentation in context', *Cognition* 127(3): 439–453, https://doi.org/10.1016/j.cognition.2013.02.002.

Labov, W. (2007). 'Transmission and diffusion', *Language* 83(2): 344–387, https://doi.org/10.1353/lan.2007.0082.

Lavi-Rotbain, O. and Arnon, I. (2022). 'The learnability consequences of Zipfian distributions in language', *Cognition* 223(August): 105038, https://doi.org/10.1016/j.cognition.2022.105038.

Leufkens, S. (2020). 'A functionalist typology of redundancy', *Revista Da Abralin* 19(3): 79–103.

Little, H., Eryılmaz, K., and De Boer, B. (2017). 'Signal dimensionality and the emergence of combinatorial structure', *Cognition* 168: 1–15.

Maldonado, M. and Culbertson, J. (2020). 'Person of interest: Experimental investigations into the learnability of person systems', *Linguistic Inquiry* 53(2): 295–336, https://doi.org/10.1162/ling_a_00406.

Maldonado, M., Culbertson, J., and Uegaki, W. (2022). 'Learnability and constraints on the semantics of clause-embedding predicates', in *Proceedings of the 44th Annual Meeting of the Cognitive Science Society*, 237–243.

Mansfield, J., Saldana, C., Hurst, P., et al. (2022). 'Category clustering and morphological learning', *Cognitive Science* 46(2): e13107, https://doi.org/10.1111/cogs.13107.

Martin, A. and Culbertson, J. (2020). 'Revisiting the suffixing preference: Native-language affixation patterns influence perception of sequences', *Psychological Science* 31(9): 1107–1116, https://doi.org/10.1177/0956797620931108.

Martin, A. and Peperkamp, S. (2020). 'Phonetically natural rules benefit from a learning bias: A re-examination of vowel harmony and disharmony', *Phonology* 37(1): 65–90, https://doi.org/10.1017/S0952675720000044.

9780192886491_Book.indb 38 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:22

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Martin, A. and White, J. (2021). 'Vowel harmony and disharmony are not equivalent in learning', *Linguistic Inquiry* 52(1): 227–239.

Martin, A., Holtz, A., Abels, K., et al. (2020). 'Experimental evidence for the influence of structure and meaning on linear order in the noun phrase', *Glossa* 5(1), https://doi.org/10.5334/GJGL.1085.

Meisezahl, M., Kirby, S., and Culbertson, J. (in press). 'Variability and learning in language change: The case of V2', *Journal of Historical Syntax*.

Merkx, M., Rastle, K., and Davis, M. H. (2011). 'The acquisition of morphological knowledge investigated through artificial language learning', *Quarterly Journal of Experimental Psychology* 64(6): 1200–1220.

Mintz, T. H., Walker, R. L., Welday, A., and Kidd, C. (2018). 'Infants' sensitivity to vowel harmony and its role in segmenting speech', *Cognition* 171: 95–107, https://doi.org/10.1016/j.cognition.2017.10.020.

Moreton, E. and Pater, J. (2012). 'Structure and substance in artificial-phonology learning, Part I: Structure', *Linguistics and Language Compass* 6(11): 702–718, https://doi.org/10.1002/lnc3.366.

Moreton, E., Pater, J., and Pertsova, K. (2015). 'Phonological concept learning', *Cognitive Science* 41(1): 4–69, https://doi.org/10.1111/cogs.12319.

Motamedi, Y., Wolters, L., Naegeli, D., et al. (2022). 'From improvisation to learning: How naturalness and systematicity shape language evolution', *Cognition* 228: 105206.

Nielsen, M., Haun, D., Kärtner, J., and Legare, C. H. (2017). 'The persistent sampling bias in developmental psychology: A call to action', *Journal of Experimental Child Psychology* 162: 31–38, https://doi.org/10.1016/j.jecp.2017.04.017.

Perfors, A. (2012). 'When do memory limitations lead to regularization? An experimental and computational investigation', *Journal of Memory and Language* 67(4): 486–506, https://doi.org/10.1016/j.jml.2012.07.009.

Raviv, L. and Arnon, I. (2018). 'Systematicity, but not compositionality: Examining the emergence of linguistic structure in children and adults using iterated learning', *Cognition* 181(August): 160–173, https://doi.org/10.1016/j.cognition.2018.08.011.

Reber, A. S. (1967). 'Implicit learning of artificial grammars', *Journal of Verbal Learning and Verbal Behavior* 6(6): 855–863.

Reber, A. S. (1989). 'Implicit learning and tacit knowledge', *Journal of Experimental Psychology: General* 118(3): 219.

Rose, S. and Walker, R. L. (2011). 'Harmony systems', in J. Goldsmith, J. Riggle, and A. C. L. Yu (eds.), *The Handbook of Phonological Theory*, 240–290. Oxford: Wiley-Blackwell.

Saffran, J. R. (2020). 'Statistical language learning in infancy', *Child Development Perspectives* 14(1): 49–54, https://doi.org/10.1111/cdep.12355.

Saffran, J. R. and Thiessen, E. D. (2003). 'Pattern induction by infant language learners', *Developmental Psychology* 39(3): 484–494, https://doi.org/10.1037/0012-1649.39.3.484.

Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996). 'Statistical learning by 8-month-old infants', *Science (New York, N.Y.)* 274(5294): 1926–1928, https://doi.org/10.1126/science.274.5294.1926.

Saldana, C., Oseki, Y., and Culbertson, J. (2021). 'Cross-linguistic patterns of morpheme order reflect cognitive biases: An experimental study of case and number morphology', *Journal of Memory and Language* 118(February): 104204, https://doi.org/10.1016/j.jml.2020.104204.

Saldana, C., Smith, K., Kirby, S., and Culbertson, J. (2021). 'Is regularization uniform across linguistic levels? Comparing learning and production of unconditioned probabilistic variation

9780192886491_Book.indb 39 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:22

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

in morphology and word order', *Language Learning and Development,* 1–31, https://doi.org/10.1080/15475441.2021.1876697.

Saldana, C., Herce, B., and Bickel, B. (2022). 'A naturalness gradient shapes the learnability and cross-linguistic distribution of morphological paradigms', in *Proceedings of the 44th Annual Meeting of the Cognitive Science Society*, 787–794.

Samara, A., Smith, K., Brown, H., and Wonnacott, E. (2017). 'Acquiring variation in an artificial language: Children and adults are sensitive to socially conditioned linguistic variation', *Cognitive Psychology* 94: 85–114, https://doi.org/10.1016/j.cogpsych.2017.02.004.

Saratsli, D., Bartell, S., and Papafragou, A. (2020). 'Cross-linguistic frequency and the learnability of semantics: Artificial language learning studies of evidentiality', *Cognition* 197: 104194, https://doi.org/10.1016/j.cognition.2020.104194.

Schwab, J. F., Lew-Williams, C., and Goldberg, A. E. (2018). 'When regularization gets it wrong: Children over-simplify language input only in production', *Journal of Child Language* 45(5): 1054–1072, https://doi.org/10.1017/S0305000918000041.

Seyfarth, S., Ackerman, F., and Malouf, R. (2014). 'Implicative organization facilitates morphological learning', in *Proceedings of the 40th Annual Meeting of the Berkeley Linguistics Society*, 480–494.

Siegelman, N., Bogaerts, L., Elazar, A., et al. (2018). 'Linguistic entrenchment: Prior knowledge impacts statistical learning performance', *Cognition* 177: 198–213, https://doi.org/10.1016/j.cognition.2018.04.011.

Skarabela, B., Ota, M., O'Connor, R., and Arnon, I. (2021). '"Clap your hands" or "take your hands"? One-year-olds distinguish between frequent and infrequent multiword phrases', *Cognition* 211: 104612, https://doi.org/10.1016/j.cognition.2021.104612.

Smith, K. (2022). 'How language learning and language use create linguistic structure', *Current Directions in Psychological Science* 31(2): 177–186, https://doi.org/10.1177/09637214211068127.

Smith, K. and Wonnacott, E. (2010). 'Eliminating unpredictable variation through iterated learning', *Cognition* 116(3): 444–449, https://doi.org/10.1016/j.cognition.2010.06.004.

Tal, S. and Arnon, I. (2020). 'Do children preferentially mark unpredictable material? The case of optional plural marking', in *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*, 3226–3232.

Tal, S. and Arnon, I. (2022). 'Redundancy can benefit learning: Evidence from word order and case marking', *Cognition* 224(February): 105055, https://doi.org/10.1016/j.cognition.2022.105055.

Tamariz, M. and Kirby, S. (2016). 'The cultural evolution of language', *Current Opinion in Psychology* 8: 37–43, https://doi.org/10.1016/j.copsyc.2015.09.003.

Tang, K., and Baer-Henney, D. (2023). 'Modelling L1 and the artificial language during artificial language learning', *Laboratory Phonology* 14(1).

Wang, F., Kirby, S., and Culbertson, J. (in press). 'The learning bias for cross-category harmony is sensitive to semantic similarity: Evidence from artificial language learning experiments', *Language*.

White, J. (2014). 'Evidence for a learning bias against saltatory phonological alternations', *Cognition* 130(1): 96–115, https://doi.org/10.1016/j.cognition.2013.09.008.

White, J. and Sundara, M. (2014). 'Biased generalization of newly learned phonological alternations by 12-month-old infants', *Cognition* 133(1): 85–90, https://doi.org/10.1016/j.cognition.2014.05.020.

9780192886491_Book.indb 40 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:22

CHAPTER 3

COMMUNICATION EXPERIMENTS

Social interaction in the formation of novel communication systems

THOMAS F. MÜLLER AND LIMOR RAVIV

Language is used for communication, i.e. people converse with each other to transmit information (Shannon and Weaver 1949) and exchange their cognitive states with others (e.g. Sperber and Wilson 1994). For example, in a day-to-day conversational situation between two people at a dinner table, one interlocutor might say, 'pass me the salt'. Their conversation partner would then pick the salt shaker from the various objects they can see on the table and pass it over. For this communicative interaction to function, interlocutors should be aligned on shared representations (Garrod and Pickering 2004) situated in their concurrent conversational context (e.g. 'the salt' referring to the shaker, not individual grains of salt; and to the shaker on our dinner table, as opposed to our neighbours' table in a restaurant; Clark 1996). These individual interactions are a basic unit in which language is used and, consequently, changes: interlocutors need to establish novel shared conventions (defined as arbitrary solutions to repeated coordination problems; Lewis 1969) for newly emerging concepts, and they can modify existing conventions in novel ways (Höfler and Smith 2009; Smith and Höfler 2015). From these dyadic micro-level changes, macro-level cultural change can arise via diffusion within a language community and via transmission to new learners (see also Chapter 1, this volume). But how do local conventions form to begin with? What pressures and processes operate during communication, and how do they shape emergent languages?

The empirical investigation of human communication and its emergence is a fruitful venue for language evolution research. In this chapter, we summarize the methodological approach of using *communication experiments* to study language evolution. We define a communication experiment as any experimental study involving two or more human participants, whose success in the task depends on signalling some crucial

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

information. This means we will not consider work on animal communication (see Part III in this volume) or computer simulations (see Part II in this volume) here. We start by providing a brief history of communication experiments (section 3.1), acknowledging classic experiments on natural language use. After this, we outline the most prominent methodological variations within communication experiments focusing on language evolution (section 3.2), including a more detailed description of two case studies focusing on the emergence of conventions within shared context and on the effect of social structure on the emergence of grammar. Lastly, we highlight the potential challenges and pitfalls of the method, focusing on participants' existing prior knowledge and the lack of data from children (section 3.3).

3.1 A SHORT HISTORY OF COMMUNICATION EXPERIMENTS

There is a long tradition of using communication experiments to explore language use, coordination, conventionalization, and—more recently—evolution. In the most standard and simple version of these experiments (see also Lewis game, Chapter 17, this volume), there is some set of stimuli (the referents) that participants need to communicate about in order to succeed. These setups are typically known as *referential communication games* or *director–matcher experiments*. In their most typical version (Figure 3.1), a pair of participants communicates about a target meaning, with one person being the sender (also referred to as the director or the producer), and the other participant being the receiver (also referred to as the matcher or the guesser). The sender needs to convey this target meaning (e.g. shape, location on a grid, etc.) in some way (e.g. using their native language, pantomime, artificial language, etc.) to the receiver, and the receiver needs to select the correct target (typically from a set of distractors).

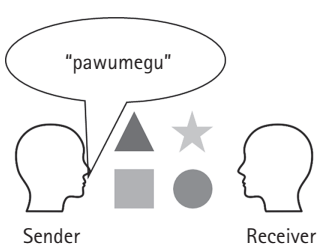


FIGURE 3.1 A basic referential communication task.
The sender needs to convey a target meaning (e.g. the triangle) to the receiver in some way (here, with a pseudoword).

When participants solve the task using natural language, the respective studies can be considered part of a field known as *experimental pragmatics* (for an overview, see e.g. Noveck 2018). Research in experimental pragmatics can tease apart the role of pragmatic

9780192886491_Book.indb 42 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:23

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

mechanisms contributing to language use by controlling and manipulating different features of the task, such as feedback, the task environment, and more. A very early example of the basic referential communication task is the series of experiments by Krauss and Weinheimer (1964, 1966, 1967), which showed that phrases used by senders to refer to visual stimuli tend to get shorter over the course of repeated conversations (Krauss and Weinheimer 1964), when receiving direct feedback from receivers or the experimenter (1966), and when stimuli are more dissimilar (1967). Clark and colleagues ran a similar task on participants' descriptions of unfamiliar *tangram* figures. These studies showed how people minimize communicative effort to—again—produce shorter phrases (Clark and Wilkes-Gibbs 1986), such that people overhearing the descriptions (but not directly participating) have a hard time understanding them (Schober and Clark 1989; Wilkes-Gibbs and Clark 1992). More recently, such paradigms have been used to investigate multimodality and the alignment of not only speech, but also gesture and manual expressions (Rasenberg, Pouw, et al. 2022; Rasenberg, Özyürek, et al. 2022).

Another classic task was created by Garrod and Anderson (1987) in the form of the 'maze game', in which participants converse in natural language to coordinate on opening barriers within a virtual maze. Results from this task emphasized the creation of local conventions between interacting participants (Garrod and Anderson 1987), e.g. by aligning on references that describe the maze via counting rows and columns. Garrod and Doherty (1994) extended this paradigm and showed that a closed community of participants will converge on shared global conventions, while isolated pairs and chains with new participants will remain relatively flexible (but also less coordinated) in their descriptions. In a recent application of the task, Nölle et al. (2020) varied the structures of the mazes and showed that linguistic adaptations were also affected by the task environment.

These examples of communication experiments already touch upon language evolution by showing how language use adapts over the course of conversation, and their novel stimuli require participants to develop new shared references. However, their reliance on natural language prevents these studies from investigating how communication can emerge in the first place, and new references developing within the system of an existing language are already greatly supported and biased by that language. The same issue applies to a related paradigm asking participants to learn novel words (together or separately) before communicating (Wu and Keysar 2007; Heller et al. 2012).

While experimental pragmatics can tell us how existing language adapts and functions in different scenarios, *experimental semiotics* makes use of communication experiments where participants are not allowed to use their native language, and instead are required to develop novel signalling systems (e.g. using gestures, drawings, vocalizations; see section 3.2.1 below) in order to communicate. Since this key innovation can inform us more clearly about language evolution, we will focus on experimental semiotics from now on. Experimental semiotics explores questions related to the formation of communication systems and is concerned with (a) how and why communication and novel conventions emerge and evolve, as well as (b) what shape and structure new conventions take. Examples for (a) include studies examining how the

9780192886491_Book.indb 43 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:23

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

emergence of conventions is affected by age (Bohn et al. 2019; Kempe et al. 2019), modality (Macuch Silva et al. 2020), or the role of shared context for emergent communication (Winters et al. 2018; Müller et al. 2019). Studies concerned with (b) investigate, for example, the form of signals when participants can repair misunderstandings (Garrod et al. 2007; Healey et al. 2007; Fay et al. 2018a) or under which social conditions people develop different types of signalling systems (Kirby et al. 2015; Motamedi et al. 2019; Raviv et al. 2019, 2020; Segovia-Martín et al. 2020). While an exhaustive review of the field of experimental semiotics is outside the scope of the current chapter (see e.g. Scott-Phillips and Kirby 2010; Galantucci and Garrod 2011; Galantucci et al. 2012; Tamariz 2017; Nölle et al. 2022; Delliponti et al. 2023), below we characterize some of the main themes within the field of experimental semiotics, focusing on methodological choices and experimental design.

3.2 Methodological choices in communication experiments

In the standard procedure of referential communication experiments, two or more participants need to interact repeatedly over several rounds of experimental trials, either face-to-face or via computer interface. Two key features of the task are the *meaning space* (i.e. what people need to communicate about) and the *signalling space* (i.e. in which form participants can communicate). During the communication experiment, participants need to map signals to meanings in a way that allows for mutual understanding. In this section, we outline many of the possible methodological variations of communication experiments in design, stimuli, and procedure, while also explaining the implications of varying these features.

3.2.1 Signalling spaces

Experimental semiotics makes participants communicate by using signals that are not *a priori* established, allowing researchers to study the formation and evolution of novel conventions instead of the use or change of established ones. The *signalling spaces* used in communication experiments can be extremely diverse and innovative and situated within different modalities. For example, participants might be asked to communicate via written letters and pseudowords (Selten and Warglien 2007; Roberts 2010; Kirby et al. 2015; Carr et al. 2017; Tamariz et al. 2018; Raviv et al. 2019), silent gestures (Christensen et al. 2016; Nölle et al. 2018; Motamedi et al. 2019; see also Chapter 9, this volume), free drawing (Galantucci 2005; Garrod et al. 2007; Healey et al. 2007; Fay et al. 2010; Galantucci et al. 2010; Fay et al. 2018a, b), pre-constructed graphical signs (Morin et al. 2022; Müller et al. 2019, 2021; Murthy et al. 2022), vocalizations (Perlman et al.

9780192886491_Book.indb 44 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:23

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

2015; Perlman and Lupyan 2018), the pitch of auditory signals (Verhoef et al. 2015; Little et al. 2017), virtual movements of participants (Newman-Norlund et al. 2009; Scott-Phillips et al. 2009; de Ruiter et al. 2010; Müller et al. 2023), or even tactile stimulations (Iizuka et al. 2013).

Across modalities, signals can be discrete (as in the case of the use of pre-existing alphabets to construct pseudowords) or continuous (as in the case of silent gestures and auditory pitch). In addition, the signalling space might be constrained in some way: for example, Raviv et al. (2019) only allowed participants to create pseudowords using ten consonants and five vowels, and Scott-Phillips et al. (2009) only allowed participants to move horizontally and vertically, but not diagonally. Other studies have left the signalling space wide open (Garrod et al. 2007; Healey et al. 2007; Fay et al. 2018a), for instance by letting participants draw freely on a whiteboard. Limiting the signalling space can have important implications for communication: for instance, Selten and Warglien (2007) found that smaller signal repertoires made emergent communication harder. It is important to design signalling spaces in line with task demands, and make sure the space is sufficiently expressive. A few studies have specifically investigated the effect of signalling modality by comparing the use of vocalizations against gestures (or both) within the same experiment (Fay et al. 2013, 2022; Macuch Silva et al. 2020; Lister et al. 2021). Results from such studies typically show that communicative success is higher for gestures compared to vocalizations, often thanks to the high degree of iconicity in the manual modality. This has potential implications for the ongoing debate about the origins of language being gestural, vocal, or multimodal.

One important design choice is whether, prior to communication, novel signals are first provided and taught to participants in an individual training regime, or whether participants are not given any initial input and are free to explore a signalling space as they please. In the case of pre-training, participants are taught predetermined signal-to-meaning mappings (which can be completely random or already carry some systematic structure), which they need to memorize during multiple rounds of exposure before they can interact with their partner(s). In the case of no pre-training, participants start communicating immediately and need to construct their first signals from scratch. The former approach can be useful if the focus is on the changes to a controlled signalling system introduced by communication (e.g. Kirby et al. 2015) or individual differences (e.g. Josserand et al. 2024), while the latter approach has been used to study the emergence and conventionalization of novel signals from scratch (Galantucci 2005; Garrod et al. 2007; Motamedi et al. 2019; Müller et al. 2019; Raviv et al. 2019).

3.2.2 Meaning spaces

The *meaning spaces* used in communication experiments (i.e. the set of referents that participants need to communicate about) also vary greatly across studies. Some studies use auditory stimuli such as music pieces (Healey et al. 2007) or environmental sounds (Macuch Silva et al. 2020), while others have used written words denoting

9780192886491_Book.indb 45 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:23

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

familiar concepts (Garrod et al. 2007; Fay et al. 2010; Perlman et al. 2015; Fay et al. 2018a; Motamedi et al. 2019; Murthy et al. 2022). However, visual stimuli are by far the most commonly used meanings. Stimuli can vary in their degree of complexity: from communicating about familiar shapes such as squares and circles (e.g. Iizuka et al. 2013) or different colours (Roberts et al. 2015; Müller et al. 2019, 2021; Morin et al. 2022), to communicating about the positions and orientation of known shapes (e.g. Newman-Norlund et al. 2009; de Ruiter et al. 2010), to communicating about more complex unknown images (e.g. Raviv et al. 2019; Macuch Silva et al. 2020), cartoon figures (Nölle et al. 2018), and even full visual scenes depicting multiple agents performing actions (Schouwstra and De Swart 2014; Christensen et al. 2016; Schouwstra 2017). The latter can be a static image or even a dynamic video depicting movement and actions.

Carefully designed meaning spaces—together with appropriate signalling spaces—have the potential to inform us about the origins of structure in natural language. Oftentimes, meaning spaces are structured to some extent, containing different combinations of features (semantic dimensions) such as shape, size, colour, motion, texture, etc., or internal categories such as professions, places, and objects (e.g. Kirby et al. 2015; Nölle et al. 2018; Tamariz et al. 2018). The idea behind using structured meaning spaces is that participants' emerging communication systems should map to fit this internal structure by creating signals that express different categories (e.g. having a part-word for describing the shape, and a part-word for describing colour). For example, Kirby et al. (2015) used a meaning space containing four discrete shapes appearing in four different fill patterns, and Nölle et al. (2018) used a meaning space containing cartoon characters that could be classified into categories based on gender, profession, and whether or not they own a pet. Meaning spaces can also be composed of combinations of more abstract semantic dimensions such as themes and temporal concepts (Motamedi et al. 2019; Verhoef et al. 2022). For example, Verhoef et al. (2022) asked participants to describe meanings representing different temporal periods ('day' vs. 'year') along three different categories (duration, sequence, deictic), and Motamedi et al. (2019) used a meaning space containing six different themes (e.g. food, religion, music, photography) that could appear in four different categories (people, locations, objects, and actions). While most meaning spaces contain discrete concepts or combinations thereof, stimuli can also be continuous: for example, Carr et al. (2017) used different types of triangles generated by varying the location of three points on a grid to create a continuous, open-ended meaning space.

Communication experiments that do not merely ask for referential communication, but are *coordination games* instead (see Galantucci and Garrod 2011), pose a particular challenge regarding the meanings in the task. Here, participants have to not only find mappings between the signals and meanings, but also agree on what common meanings they want to refer to in the first place. In other words, while success in a referential experiment is achieved by finding shared signals for a pre-defined referent (the 'target'), in a coordination game, participants have to reach a common goal by finding useful referents plus the signals for them; this is closer to the semiotic challenge encountered in real-life communication. One example of this is experiments based on Galantucci

9780192886491_Book.indb 46 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:23

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

(2005), in which the participants' goal is to move to the same room in a virtual space; but to do so, they also have to agree on what exactly should be communicated about locations. For instance, some pairs in Galantucci (2005) converged on signals that numbered the different rooms in the space, some pairs referred to visual icons associated with different rooms, and other pairs described the longitudinal location of the player in the virtual space.

Some experiments have made the impact of the meaning space their object of investigation and systematically manipulated it, focusing on emergent structure in the signals. For example, Selten and Warglien (2007) showed that a growing meaning space led to the evolution of more compositional structure in the signals to meet the demands of communicating new referents. Similarly, Nölle et al. (2018) found that a more open (i.e. not limited to a small set of referents) meaning space led to higher systematicity in gestural signals. They also showed that functionally adaptive structure in the meaning space (i.e. what feature of a referent mattered for accurate discrimination) was reflected in emergent signal structure. A few studies also considered how particular meanings can afford iconic signals to develop, affecting internal signal order (Christensen et al. 2016) or preventing the emergence of combinatorial signals (Roberts et al. 2015). Lastly, more frequently occurring meanings have shown to lead to shorter signals when there is pressure to communicate accurately and efficiently (Kanwal et al. 2017), and signal order tends to conventionalize towards the most frequent order in the meaning space (Christensen et al. 2016). Taken together, these examples illustrate that it is important to choose a particular meaning space in line with task demands, as it will impact communication and the resulting signals majorly. For our understanding of language evolution, they show how language can be shaped by the meanings that interlocutors are trying to communicate.

3.2.3 Feedback and repair

Communication experiments also vary in the amount of *feedback* that participants receive. We use this term to specifically refer to information received periodically after a trial in the task has ended. Providing full feedback would mean that all participants receive information not only about success or failure, but also about what choice was made regarding the referents (e.g. Raviv et al. 2019, 2020). Partial feedback can leave out this information on choices (e.g. Healey et al. 2007; Moreno and Baggio 2015), limit information to only one participant, or even obfuscate the feedback on performance by only providing the average performance over a block of trials (e.g. Müller et al. 2021; Morin et al. 2022). Feedback is a straightforward design choice to calibrate the difficulty of a task, but not many communication experiments have investigated its impact on language evolution directly. A notable exception is the study by de Ruiter et al. (2010), which ruled out that senders are merely 'dictating' encoding strategies to receivers by showing that communication improved when feedback was available. Importantly, providing extensive feedback can change the interpretation of a communication experiment because

9780192886491_Book.indb 47 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:23

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

signal–meaning associations can arise from repeated trial and error reinforced by the feedback. If a study is concerned with the emergence of communication from scratch, this can be a potential confound, which has led some studies to forgo feedback completely (e.g. Iizuka et al. 2013; Fay et al. 2018a; Müller et al. 2019).

A design choice related to feedback is whether participants in the task are able to *repair* misunderstandings that might arise, which is a ubiquitous phenomenon in natural language (Dingemanse et al. 2015). The term is sometimes used interchangeably with 'feedback' in the literature, but we want to keep the two clearly separate. The key difference for us is that repair gives participants themselves the option to indicate (mis) understanding (for example, think of interjections like 'sorry?' or 'uh-huh' in English) and potentially mend it, whereas feedback is provided by the experimental design (note that a similar distinction was already made by Krauss and Weinheimer in 1966, albeit in a task involving natural language). Implementation-wise, communication experiments can give pre-defined repair signals to participants (Müller et al. 2019) or they can simply let them interact freely with one another in games with open signalling spaces (Garrod et al. 2007; Healey et al. 2007; Fay et al. 2018a) to, for example, 'tick' solutions they understood. Repair can be a helpful option for participants while adding to the ecological validity of the task. However, previous research focused on repair by itself (Garrod et al. 2007; Healey et al. 2007; Fay et al. 2018a) has only found minor improvements in communicative success, while showing benefits for developing more abstract signals, i.e. signals that are less iconic, more simplified, and harder to understand for outsiders. Garrod et al. (2007) also showed that this symbolification happened faster when repair was concurrent, as opposed to after drawing had already finished.

3.2.4 Sender and receiver roles

Regardless of the amount and manner of feedback or repair, the basic interactions between participants can also differ fundamentally in the way *sender/receiver roles* are assigned. A typical design is to only provide information about the correct referent to one participant at a time and ask them to produce signals, while the other participant tries to understand and choose. To keep these roles controlled, they can either stay the same for the entirety of the experiment, which means communication is inherently asymmetric and the receiver is limited to a reactive role; or participants can alternate and switch roles. The latter is arguably more interactive and realistic given the turn-taking structure of natural language (Sacks et al. 1974), but the former can be useful to keep the challenges of sending/receiving separate. In fact, role asymmetry can be beneficial for language evolution because it means only one side has to coordinate if a task can be solved by 'dictating' an arbitrary code to the receiver (Selten and Warglien 2007; Moreno and Baggio 2015). Alternatively, both participants can also be tasked with signalling and understanding at the same time. This is frequently the case in coordination games, where oftentimes the turn-taking structure itself (or, in fact, what constitutes a turn) is not even defined, but has to be organized by the participants

9780192886491_Book.indb 48 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:23

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

themselves (e.g. Galantucci 2005; Scott-Phillips et al. 2009). Such a design increases the ecological validity, but also the communicative challenge of the task. A few studies have also considered the performance of 'overhearers', i.e. participants that did not participate in the original interactions of an experiment. Overhearers end up with less accurate, less efficient, and less aligned communication (Fay et al. 2018a), and performance is particularly bad when they cannot observe signals from the early stages of communication (Garrod et al. 2007). However, iconicity can help overhearers with understanding the original signals (Perlman et al. 2015; Perlman and Lupyan 2018). Ultimately, investigating the roles that participants take on in communication experiments can inform us about what makes a resulting 'language' shared.

3.2.5 Case study 1: Communicative contexts

Referents are usually shown in contrast with other options to pose a communicative challenge. As a real-life analogue, imagine standing in front of a shelf at the supermarket while being told to 'get the orange juice', which is next to other juice boxes, bottles of water, etc. These communicative *contexts* are dependent on the meaning space, but they can be varied separately even within one constant meaning space. A straightforward contextual variation is that increasing the number of referents can make communication about these referents harder. Notably, this affects both the sender (who has to signal more precisely to denote a referent from a larger context) and the receiver (who has to pick from a larger context) in the experiment. But also the distance between referents can affect the difficulty in a similar vein (e.g. more similar vs. less similar colours). A number of communication experiments have manipulated the communicative context to show how it affects language evolution. Similar to the results for meaning spaces, some studies showed that the relevant dimensions within contexts will also change how signals get structured (e.g. Winters et al. 2015, 2018). Murthy et al. (2022) manipulated the variability of participants' prior associations within contexts, showing that contexts with well-aligned priors are easier to coordinate on. Manipulating contexts also allows for varying the information that is shared between participants in any given trial. In line with classic theories of communication (e.g. Grice 1989; Sperber and Wilson 1994; Clark 1996), more shared information between participants is associated with better performance during communication (de Ruiter et al. 2010; Winters et al. 2018; Müller et al. 2019).

We present the study by Müller et al. (2019) in more detail here. This study used the same referential communication task in two experiments to investigate the impact of shared information on emergent communication. Participants were asked to communicate the correct colour from an array of four colours by using pre-constructed graphic symbols. These symbols were black-and-white and chosen to be ambiguous regarding their association to the colour space in the task (Figure 3.2a). Participants were enabled to repair potential misunderstandings via pre-defined written annotations that were introduced to them before the experiment, but received no feedback from

9780192886491_Book.indb 49 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:23

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

the experimental design itself. The shared visual context between participants in both experiments was manipulated by restricting the sender's access to only one colour (the target) from the array the receiver had to choose from (Figure 3.2b). The results of both experiments showed that participants' emergent communication benefitted from the senders' access to the context. Performance also improved over time, suggesting participants were gradually establishing local conventions. Moreover, switching to a different meaning space in the second half of Experiment 2 revealed that participant pairs benefitted from reusing their existing conventions only when they had had access to the shared context before, suggesting that those communication systems were more flexible.

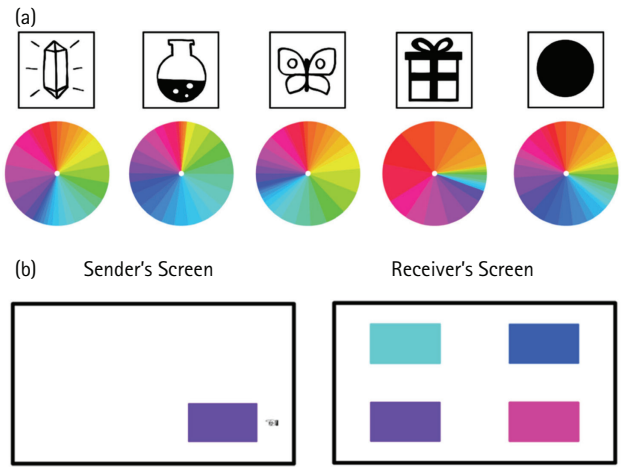


FIGURE 3.2 Case study on shared visual context.

Panel (a) shows five examples of black-and-white symbols that participants could use to communicate about different colours, along with their corresponding mapping to colours as reported by senders in Experiment 2 in Müller et al. (2019). The area occupied by a colour corresponds to the number of reported uses for that colour. The figure shows large variance across senders, suggesting sufficient ambiguity of the black-and-white symbols; but also patterns specific to individual symbols, like a wide blue-green association for the second symbol or a concentration on the red area for the fourth one. Panel (b) shows an example screen from the 'unshared visual context condition'. In the 'shared' condition, the sender would see all four colours instead, arranged randomly.

Source: adapted with permission from Müller et al. (2019); reproduced under a CC BY-NC license.

3.2.6 Case study 2: Social structure

Communication experiments can also be used to study the relationship between language evolution and the social environment in which languages evolve. This is done by manipulating the number of interacting participants (e.g. Fay et al. 2008, 2010; Raviv et al. 2019) and/or the degree of connectivity between participants (e.g. Raviv et al. 2020; Segovia-Martín et al. 2020). Since the classic communication experiment is dyadic in nature (i.e. featuring a sender and a receiver), social structure variations are typically implemented by splitting groups into pairs and having participants switch

9780192886491_Book.indb 50 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:24

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

partners from time to time. While one-to-many interactions are easy to implement (e.g. one participant sends a signal to multiple members of the group, who need to guess it), they introduce an asymmetry in production and comprehension that is often unwanted. They can also substantially speed up the process of convergence and thus hinder our ability to see how individual interactions build up to population-level phenomena.

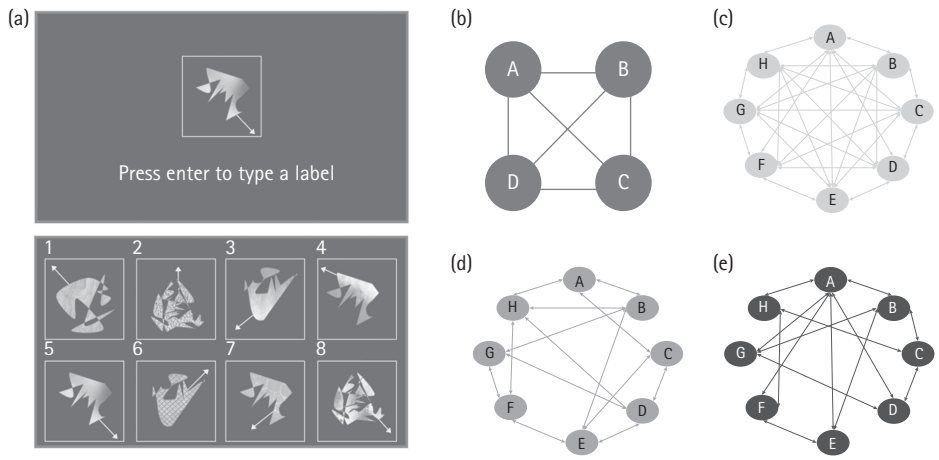


FIGURE 3.3 Case study on the role of social structure using a group communication paradigm.

Panel (a) shows the sender's and receiver's screens, respectively. Raviv et al. (2019) tested the effect of community size by comparing fully connected groups of four (b) or eight (c) participants, and Raviv et al. (2020) tested the role of network structure by comparing groups of eight participants that were either fully connected (c), small world (d), or scale free (e).

Source: Adapted from Raviv et al. (2020); reproduced under a CC BY license.

Zooming in on the role of the social environment, Raviv et al. (2019) tested the effect of community size in the formation of artificial languages using the 'group communication game'. In the study, different-sized groups needed to use written pseudowords to communicate about a set of novel scenes, which combined discrete shapes with continuous motion by having shapes move in an angle from a 360 degree range (Figure 3.3). Groups were composed of either four or eight participants in order to evaluate whether bigger groups would create more structured languages—a prediction motivated by cross-linguistic studies and theories of language change which suggest that languages spoken in big communities tend to have more systematic and transparent grammars (Wray and Grace 2007; Trudgill 2008, 2009; Lupyan and Dale 2010). The results of this study showed that larger groups developed more compositional languages over time, and did so faster and more consistently than small groups. Larger groups were seemingly under a stronger pressure to generalize their languages and favour systematic variants, which could, in turn, ease mutual understanding and facilitate convergence. Moreover, results showed that small groups varied more in their linguistic behaviours, while larger groups behaved relatively similarly to each other. Together, the results of Raviv et al. (2019) show that community size had a unique and causal influence on the formation of new languages.

9780192886491_Book.indb 51 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:24

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

A similar design was used in Raviv et al. (2020) to test the role of network structure, that is, what happens when groups are of the same size, but their degree of density and connectivity is manipulated. Specifically, Raviv et al. (2020) compared communication in fully connected networks (in which all participants interact with each other), small-world networks (where some participants never directly interact, but are still indirectly linked by a short chain of shared connections), and scale-free networks (which are similar to small-world networks in terms of sparsity, but where the distribution of connections follows a power law: such that most participants only interact with a few people, yet some—the 'hubs'—interact with nearly everyone in the group). Results showed that in such relatively small groups, the exact architecture of the network did not significantly affect the evolution of new languages: all networks developed languages that were highly systematic and compositional, communicatively efficient, stable, and shared across members. However, small-world networks showed the most variance in behaviour—suggesting that such networks may be more sensitive to random events (i.e. drift), potentially because frequent interactions among small subgroups can preserve random behaviours more easily.

Finally, communication experiments can also be combined with iterated learning paradigms (see Chapter 1, this volume), incorporating both interaction and transmission across multiple 'generations' of communication pairs (e.g. Kirby et al. 2015; Motamedi et al. 2019) or groups. In such setups, the final communication systems created by the first dyad or group (i.e. their signal-to-meaning mappings in the last communication round) become the input for the next dyad or group in the chain, and so on for multiple iterations. This can help disentangle the effects of transmission from the effects of communication and show the unique contribution of each process.

3.2.7 Practical considerations and measures

We have alluded to the communicative challenge of the task several times already when discussing individual methodological choices, but task difficulty ultimately is a result of the combination of all these features, interacting with one another. If communication in the experiment turns out to be trivial or impossible to achieve, this can mask important effects, as differences between participants might be hard to obtain. However, failures to communicate can also be insightful and an indication that a feature that is core to the emergence of communication might be missing from the task, or from the approach that participants are taking to it (Galantucci and Garrod 2011).

Communication experiments can be either face-to-face or online, with or without a computer interface (see section 3.3). The data from such experiments typically include a record of the target item, the distractor items, the selected items, matching accuracy, sender and receiver identity, trial number and round number, accompanied by a record of the signal used for communication (e.g. in written form in case of pseudowords, or using video/audio recording for gestures and vocalizations).

9780192886491_Book.indb 52 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:24

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Depending on the objectives of the study, researchers can analyse the changes to the emergent communication systems over time based on parameters such as communicative success (typically binary, indicating whether the target item was guessed correctly or not), signal length (e.g. the number of characters in a pseudoword or the duration of a gesture), iconicity of the signal (i.e. how easy it is for naive participants to guess the meaning of the signal), the degree of convergence between participants (i.e. how similar are the signals produced by different participants for the same meaning) or within participants (i.e. how stable are signals used by the same participant over time), and the degree of systematic structure and/or complexity of the signals. The latter measure is typically of great interest to language evolution researchers as it may indicate the emergence of grammar during communication, and it can be quantified in multiple ways: for example, by using topographic similarity in form–meaning mapping (e.g. Kirby et al. 2008, 2015; Nowak and Baggio 2016), conditional entropy (e.g. Winters et al. 2015), or the form combination index (e.g. Roberts and Galantucci 2012).

3.3 Common challenges for communication experiments

In this section, we focus on challenges specific to communication experiments that are beyond the general limitations of experimental work. First, communication experiments face a specific recruitment challenge given that they involve two or more participants: scheduling and coordinating between multiple participants to join the experiment at the same time can be difficult, and if a single participant does not show up or drops out during the task, the whole session is at risk. One way to handle this issue is to send multiple confirmation and reminder emails and/or call participants directly prior to the experiment (in case of in-house testing), as well as having a backup participant at every time slot who is willing to participate on short notice in case of a last-minute cancellation. Given these constraints, data collection for communication experiments (especially those involving groups) can take longer than for regular experiments. Moreover, if the experiment is held face-to-face, it often requires the presence of an experimenter in the room to closely monitor participants' behaviour and ensure they are following the instructions and not using other means to communicate.

One possible way of overcoming these challenges is to turn to online experiments. The main advantage of an online setup is the ability to recruit many more participants from various locations, and test them simultaneously without the need to bring them physically to the lab. The availability of individuals may be higher when tested from home, the chances of no-shows may be reduced, and the testing costs may be relieved. Specialized setups like a custom smartphone application (Morin et al. 2018) can even open up unique advantages such as free partner choice or testing larger groups of participants, which would not be feasible in the lab. However, online communication

9780192886491_Book.indb 53 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:24

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

experiments also introduce novel challenges. For example, participants tested remotely may be engaged in other activities during the experiment, which may affect their performance, and may be more likely to drop out during the experiment. Moreover, online experiments are less naturalistic, as they do not include actual face-to-face interaction, and time-sensitive interactions between participants might get distorted by lag.

Arguably the largest methodological challenge for experimental semiotics is that participants' prior knowledge may affect the results of communication experiments: since participants are already fluent in one or more languages before starting the experiment, this might change their behaviour. For example, participants' linguistic experience may affect the phonology or morphosyntax of a specific emerging communication system, for instance, by producing vocalizations that follow the phonotactic rules of their native language, or by favouring the word order of their native language. Similarly, Müller et al. (2021) showed that the semantic structure participants created for colour terms in an artificial language closely resembled the one employed by their language community of English, German, or French. The specific biases introduced by participants' native language are hard to assess, yet could potentially be tested by comparing the communication systems created by participants from diverse linguistic backgrounds (see Żywiczyński et al. 2021). More crucial, however, is the concern that participants' rich prior linguistic knowledge is driving the general patterns observed in communication studies, such as the emergence of grammar or alignment on shared signals. If participants are relying heavily on their prior linguistic and semiotic knowledge, it may imply that rather than helping us understand the processes and mechanisms that are *responsible* for the evolution of language, what we see in communication experiments may simply be the *result* of such processes. These points are serious concerns which could be addressed (but not entirely solved) in several ways.

First, when designing signal and meaning spaces, it is important to avoid straightforward relations to established conventions in participants' known language(s), which they would be able to rely on. This problem is, however, not limited only to languages, but can arise from any prior shared association that participants can exploit, e.g. due to other cultural knowledge (e.g. describing an inverted-triangle shape as 'red' because of the traffic sign), or unintended iconic affordances (e.g. describing abstract diagonal lines as 'green' because they resemble a tree; of course, iconicity can be the entire point, like in many studies using gestures or free drawing). Regardless of the experimental parameters, trying to make sure participants are not acquainted with each other is also important (e.g. by inviting them separately), since they might be able to build on their past shared experience. As such, during the recruitment process, it is wise to avoid recruiting participants who are already familiar with one another, and to ensure that participants do not meet before the task.

Second, extending existing findings to child participants (who have far less extensive and explicit linguistic knowledge) is an important step in validating and replicating the patterns obtained by communication experiments with adults. However, the vast majority of the literature has focused on communication experiments with adult participants only, with only a handful of studies using child-friendly paradigms

9780192886491_Book.indb 54 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:24

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

to examine communication patterns in child–child and/or in child–adult dyads (Anderson et al. 1991, 1994; Garrod and Clark 1993; Bohn et al. 2019; Kempe et al. 2019; Lister et al. 2021). The scarcity of communication experiments with children is likely due to children's generally poorer performance in artificial language paradigms (e.g. Ferman and Karni 2010; Raviv and Arnon 2018), and their poorer communicative skills (e.g. Anderson et al. 1991, 1994; Garrod and Clark 1993).

Nevertheless, a handful of studies did look at child communicators, and these can provide valuable insights into the potential similarities and differences between children and adults in terms of their potential role (or lack thereof) in the evolution of language. This is of particular interest given work on emerging sign languages (Senghas and Coppola 2001; Senghas et al. 2004) and creoles (e.g. Bickerton 1984), which suggests that children have a special role in creating core properties of language by introducing structural innovations to novel communication systems. For example, Bohn et al. (2019) and Lister et al. (2021) showed that 6-year-olds are able to spontaneously create novel gestural signs in a communicative setup in a similar way to adults, but that their performance and ability to create compositional structure increased with age. Varying participants' age could also reveal how it may affect the emergence of conventionalized signals and/or grammar. Notably, the few communication studies conducted with children provide only weak support for children's roles as agents of language evolution, and suggest that adults are the ones scaffolding the process of conventionalization and the emergence of compositionality (e.g. Newman-Norlund et al. 2009; Kempe et al. 2019).

3.4 Conclusions

In this chapter, we have outlined some of the most important methodological aspects of communication experiments in the service of language evolution research, and demonstrated how this method can be used to tackle different questions related to the emergence of communicative conventions and grammatical structure. Notably, similar paradigms can be employed with artificial agents (Chapter 17, this volume), and communication experiments can fruitfully combine with other paradigms such as iterated learning (Chapter 1, this volume) to generate novel insights into the evolution of language.

References

Anderson, A. H., Clark, A., and Mullin, J. (1991). 'Introducing information in dialogues: Forms of introduction chosen by young speakers and the responses elicited from young listeners', *Journal of Child Language* 18(3): 663–687, https://doi.org/10.1017/S0305000900011302.

Anderson, A. H., Clark, A., and Mullin, J. (1994). 'Interactive communication between children: Learning how to make language work in dialogue', *Journal of Child Language* 21(2): 439–463, https://doi.org/10.1017/S030500090000934X.

9780192886491_Book.indb 55 /12_first_proofs/first_pr/xml_for_typesetting 27-Mar-25 23:02:24

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Bickerton, D. (1984). 'The language bioprogram hypothesis', *Behavioral and Brain Sciences* 7(2): 173–188, https://doi.org/10.1017/S0140525X00044149.

Bohn, M., Kachel, G., and Tomasello, M. (2019). 'Young children spontaneously recreate core properties of language in a new modality', *Proceedings of the National Academy of Sciences* 116(51): 26072–26077, https://doi.org/10.1073/pnas.1904871116.

Carr, J. W., Smith, K., Cornish, H., and Kirby, S. (2017). 'The cultural evolution of structured languages in an open-ended, continuous world', *Cognitive Science* 41(4): 892–923, https://doi.org/10.1111/cogs.12371.

Christensen, P., Fusaroli, R., and Tylén, K. (2016). 'Environmental constraints shaping constituent order in emerging communication systems: Structural iconicity, interactive alignment and conventionalization', *Cognition* 146: 67–80, https://doi.org/10.1016/j.cognition.2015.09.004.

Clark, H. H. (1996). *Using Language*. Cambridge: Cambridge University Press.

Clark, H. H. and Wilkes-Gibbs, D. (1986). 'Referring as a collaborative process', *Cognition* 22(1): 1–39, https://doi.org/10.1016/0010-0277(86)90010-7.

de Ruiter, J. P., Noordzij, M. L., Newman-Norlund, S., et al. (2010). 'Exploring the cognitive infrastructure of communication', *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems* 11(1): 51–77, https://doi.org/10.1075/is.11.1.05rui.

Delliponti, A., Raia, R., Sanguedolce, G., et al. (2023). 'Experimental semiotics: A systematic categorization of experimental studies on the bootstrapping of communication systems', *Biosemiotics* 16(2): 291–310, https://doi.org/10.1007/s12304-023-09534-x.

Dingemanse, M., Roberts, S. G., Baranova, J., et al. (2015). 'Universal principles in the repair of communication problems', *PLOS ONE* 10(9): e0136100, https://doi.org/10.1371/journal.pone.0136100.

Fay, N., Garrod, S., and Roberts, L. (2008). 'The fitness and functionality of culturally evolved communication systems', *Philosophical Transactions of the Royal Society B: Biological Sciences* 363(1509): 3553–3561, https://doi.org/10.1098/rstb.2008.0130.

Fay, N., Garrod, S., Roberts, L., and Swoboda, N. (2010). 'The interactive evolution of human communication systems', *Cognitive Science* 34(3): 351–386, https://doi.org/10.1111/j.1551-6709.2009.01090.x.

Fay, N., Arbib, M., and Garrod, S. (2013). 'How to bootstrap a human communication system', *Cognitive Science* 37(7): 1356–1367, https://doi.org/10.1111/cogs.12048.

Fay, N., Walker, B., Swoboda, N., and Garrod, S. (2018a). 'How to create shared symbols', *Cognitive Science* 42: 241–269, https://doi.org/10.1111/cogs.12600.

Fay, N., Walker, B., Swoboda, N., et al. (2018b). 'Universal principles of human communication: Preliminary evidence from a cross-cultural communication game', *Cognitive Science* 42(7): 2397–2413, https://doi.org/10.1111/cogs.12664.

Fay, N., Walker, B., Ellison, T. M., et al. (2022). 'Gesture is the primary modality for language creation', *Proceedings of the Royal Society B: Biological Sciences* 289(1970): 20220066, https://doi.org/10.1098/rspb.2022.0066.

Ferman, S. and Karni, A. (2010). 'No childhood advantage in the acquisition of skill in using an artificial language rule', *PLoS ONE* 5(10): e13648, https://doi.org/10.1371/journal.pone.0013648.

Galantucci, B. (2005). 'An experimental study of the emergence of human communication systems', *Cognitive Science* 29(5): 737–767, https://doi.org/10.1207/s15516709cog0000_34.

Galantucci, B. and Garrod, S. (2011). 'Experimental semiotics: A review', *Frontiers in Human Neuroscience* 5: Article 11, https://doi.org/10.3389/fnhum.2011.00011.

9780192886491_Book.indb 56 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:24

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Galantucci, B., Kroos, C., and Rhodes, T. (2010). 'The effects of rapidity of fading on communication systems', *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems* 11(1): 100–111, https://doi.org/10.1075/is.11.1.03gal.

Galantucci, B., Garrod, S., and Roberts, G. (2012). 'Experimental semiotics', *Language and Linguistics Compass* 6(8): 477–493, https://doi.org/10.1002/lnc3.351.

Garrod, S. and Anderson, A. (1987). 'Saying what you mean in dialogue: A study in conceptual and semantic co-ordination', *Cognition* 27(2): 181–218, https://doi.org/10.1016/0010-0277(87)90018-7.

Garrod, S. and Clark, A. (1993). 'The development of dialogue co-ordination skills in schoolchildren', *Language and Cognitive Processes* 8(1): 101–126, https://doi.org/10.1080/01690969308406950.

Garrod, S. and Doherty, G. (1994). 'Conversation, co-ordination and convention: An empirical investigation of how groups establish linguistic conventions', *Cognition* 53(3): 181–215, https://doi.org/10.1016/0010-0277(94)90048-5.

Garrod, S. and Pickering, M. J. (2004). 'Why is conversation so easy?', *Trends in Cognitive Sciences* 8(1): 8–11, https://doi.org/10.1016/j.tics.2003.10.016.

Garrod, S., Fay, N., Lee, J., et al. (2007). 'Foundations of representation: Where might graphical symbol systems come from?', *Cognitive Science* 31(6): 961–987, https://doi.org/10.1080/03640210701703659.

Grice, P. (1989). *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

Healey, P. G., Swoboda, N., Umata, I., and King, J. (2007). 'Graphical language games: Interactional constraints on representational form', *Cognitive Science* 31(2): 285–309, https://doi.org/10.1080/15326900701221363.

Heller, D., Gorman, K. S., and Tanenhaus, M. K. (2012). 'To name or to describe: Shared knowledge affects referential form', *Topics in Cognitive Science* 4(2): 290–305, https://doi.org/10.1111/j.1756-8765.2012.01182.x.

Höfler, S. H. and Smith, A. D. M. (2009). 'The pre-linguistic basis of grammaticalisation: A unified approach to metaphor and reanalysis', *Studies in Language* 33(4): 886–909, https://doi.org/10.1075/sl.33.4.03hoe.

Iizuka, H., Marocco, D., Ando, H., and Maeda, T. (2013). 'Experimental study on co-evolution of categorical perception and communication systems in humans', *Psychological Research* 77(1): 53–63, https://doi.org/10.1007/s00426-012-0420-5.

Josserand, M., Pellegrino, F., Grosseck, O., et al. (2024). 'Adapting to individual differences: An experimental study of language evolution in heterogeneous populations', *Cognitive Science* 48(11): e70011, https://onlinelibrary.wiley.com/doi/10.1111/cogs.70011.

Kanwal, J., Smith, K., Culbertson, J., and Kirby, S. (2017). 'Zipf's Law of Abbreviation and the Principle of Least Effort: Language users optimise a miniature lexicon for efficient communication', *Cognition* 165: 45–52, https://doi.org/10.1016/j.cognition.2017.05.001.

Kempe, V., Gauvrit, N., Gibson, A., and Jamieson, M. (2019). 'Adults are more efficient in creating and transmitting novel signalling systems than children', *Journal of Language Evolution* 4(1): 44–70, https://doi.org/10.1093/jole/lzy012.

Kirby, S., Cornish, H., and Smith, K. (2008). 'Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language', *Proceedings of the National Academy of Sciences* 105(31): 10681–10686, https://doi.org/10.1073/pnas.0707835105.

Kirby, S., Tamariz, M., Cornish, H., and Smith, K. (2015). 'Compression and communication in the cultural evolution of linguistic structure', *Cognition* 141: 87–102, https://doi.org/10.1016/j.cognition.2015.03.016.

9780192886491_Book.indb 57 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:24

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Krauss, R. M. and Weinheimer, S. (1964). 'Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study', *Psychonomic Science* 1: 113–114, https://doi.org/10.3758/BF03342817.

Krauss, R. M. and Weinheimer, S. (1966). 'Concurrent feedback, confirmation, and the encoding of referents in verbal communication', *Journal of Personality and Social Psychology* 4(3): 343–346, https://doi.org/10.1037/h0023705.

Krauss, R. M. and Weinheimer, S. (1967). 'Effect of referent similarity and communication mode on verbal encoding', *Journal of Verbal Learning and Verbal Behavior* 6(3): 359–363, https://doi.org/10.1016/S0022-5371(67)80125-7.

Lewis, D. (1969). *Convention*. Cambridge, MA: Harvard University Press.

Lister, C. J., Burtenshaw, T., Walker, B., et al. (2021). 'A cross-sectional test of sign creation by children in the gesture and vocal modalities', *Child Development* 92(6): 2395–2412, https://doi.org/10.1111/cdev.13587.

Little, H., Eryilmaz, K., and De Boer, B. (2017). 'Conventionalisation and discrimination as competing pressures on continuous speech-like signals', *Interaction Studies: Social Behaviour and Communication in Biological and Artificial Systems* 18(3): 352–375, https://doi.org/10.1075/is.18.3.04lit.

Lupyan, G. and Dale, R. (2010). 'Language structure is partly determined by social structure', *PLoS ONE* 5(1): e8559, https://doi.org/10.1371/journal.pone.0008559.

Macuch Silva, V., Holler, J., Özyürek, A., and Roberts, S. G. (2020). 'Multimodality and the origin of a novel communication system in face-to-face interaction', *Royal Society Open Science* 7(1): 182056, https://doi.org/10.1098/rsos.182056.

Moreno, M. and Baggio, G. (2015). 'Role asymmetry and code transmission in signaling games: An experimental and computational investigation', *Cognitive Science* 39(5): 918–943, https://doi.org/10.1111/cogs.12191.

Morin, O., Winters, J., Müller, T. F., et al. (2018). 'What smartphone apps may contribute to language evolution research', *Journal of Language Evolution* 3(2): 91–93, https://doi.org/10.1093/jole/lzy005.

Morin, O., Müller, T. F., Morisseau, T., and Winters, J. (2022). 'Cultural evolution of precise and agreed-upon semantic conventions in a multiplayer gaming app', *Cognitive Science* 46(2): e13113, https://doi.org/10.1111/cogs.13113.

Motamedi, Y., Schouwstra, M., Smith, K., et al. (2019). 'Evolving artificial sign languages in the lab: From improvised gesture to systematic sign', *Cognition* 192: 103964, https://doi.org/10.1016/j.cognition.2019.05.001.

Müller, T. F., Winters, J., and Morin, O. (2019). 'The influence of shared visual context on the successful emergence of conventions in a referential communication task', *Cognitive Science* 43(9): e12783, https://doi.org/10.1111/cogs.12783.

Müller, T. F., Winters, J., Morisseau, T., et al. (2021). 'Colour terms: Native language semantic structure and artificial language structure formation in a large-scale online smartphone application', *Journal of Cognitive Psychology* 33(4): 357–378, https://doi.org/10.1080/20445911.2021.1900199.

Müller, T. F., Brinkmann, L., Winters, J., and Pescetelli, N. (2023). 'Machine impostors can avoid human detection and interrupt the formation of stable conventions by imitating past interactions: A minimal Turing test', *Cognitive Science* 47(4): e13288, https://doi.org/10.1111/cogs.13288.

Murthy, S. K., Griffiths, T. L., and Hawkins, R. D. (2022). 'Shades of confusion: Lexical uncertainty modulates ad hoc coordination in an interactive communication task', *Cognition* 225: 105152, https://doi.org/10.1016/j.cognition.2022.105152.

9780192886491_Book.indb 58 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:24

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Newman-Norlund, S. E., Noordzij, M. L., Newman-Norlund, R. D., et al. (2009). 'Recipient design in tacit communication', *Cognition* 111(1): 46–54, https://doi.org/10.1016/j.cognition.2008.12.004.

Nölle, J. and Galantucci, B. (2022). 'Experimental semiotics: Past, present, and future', in A. M. García and A. Ibáñez (eds.), *The Routledge Handbook of Semiosis and the Brain*, 66–81. London: Routledge.

Nölle, J., Staib, M., Fusaroli, R., and Tylén, K. (2018). 'The emergence of systematicity: How environmental and communicative factors shape a novel communication system', *Cognition* 181: 93–104, https://doi.org/10.1016/j.cognition.2018.08.014.

Nölle, J., Fusaroli, R., Mills, G. J., and Tylén, K. (2020). 'Language as shaped by the environment: Linguistic construal in a collaborative spatial task', *Palgrave Communications* 6(1): 27, https://doi.org/10.1057/s41599-020-0404-9.

Noveck, I. (2018). *Experimental Pragmatics: The Making of a Cognitive Science*. Cambridge: Cambridge University Press.

Nowak, I. and Baggio, G. (2016). 'The emergence of word order and morphology in compositional languages via multigenerational signaling games', *Journal of Language Evolution* 1(2): 137–150, https://doi.org/10.1093/jole/lzw007.

Perlman, M. and Lupyan, G. (2018). 'People can create iconic vocalizations to communicate various meanings to naïve listeners', *Scientific Reports* 8(1): 2634, https://doi.org/10.1038/s41598-018-20961-6.

Perlman, M., Dale, R., and Lupyan, G. (2015). 'Iconicity can ground the creation of vocal symbols', *Royal Society Open Science* 2(8): 150152, https://doi.org/10.1098/rsos.150152.

Rasenberg, M., Özyürek, A., Bögels, S., and Dingemanse, M. (2022). 'The primacy of multimodal alignment in converging on shared symbols for novel referents', *Discourse Processes* 59(3): 209–236, https://doi.org/10.1080/0163853X.2021.1992235.

Rasenberg, M., Pouw, W., Özyürek, A., and Dingemanse, M. (2022). 'The multimodal nature of communicative efficiency in social interaction', *Scientific Reports* 12(1): 19111, https://doi.org/10.1038/s41598-022-22883-w.

Raviv, L. and Arnon, I. (2018). 'Systematicity, but not compositionality: Examining the emergence of linguistic structure in children and adults using iterated learning', *Cognition* 181: 160–173, https://doi.org/10.1016/j.cognition.2018.08.011.

Raviv, L., Meyer, A., and Lev-Ari, S. (2019). 'Larger communities create more systematic languages', *Proceedings of the Royal Society B: Biological Sciences 286*(1907): 20191262, https://doi.org/10.1098/rspb.2019.1262.

Raviv, L., Meyer, A., and Lev-Ari, S. (2020). 'The role of social network structure in the emergence of linguistic structure', *Cognitive Science* 44(8): e12876, https://doi.org/10.1111/cogs.12876.

Roberts, G. (2010). 'An experimental study of social selection and frequency of interaction in linguistic diversity', *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems* 11(1): 138–159, https://doi.org/10.1075/is.11.1.06rob.

Roberts, G. and Galantucci, B. (2012). 'The emergence of duality of patterning: Insights from the laboratory', *Language and Cognition* 4(4): 297–318, https://doi.org/10.1515/langcog-2012-0017.

Roberts, G., Lewandowski, J., and Galantucci, B. (2015). 'How communication changes when we cannot mime the world: Experimental evidence for the effect of iconicity on combinatoriality', *Cognition* 141: 52–66, https://doi.org/10.1016/j.cognition.2015.04.001.

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). 'A simplest systematics for the organization of turn-taking for conversation', *Language* 50(4): 696–735.

9780192886491_Book.indb 59 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:25

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Schober, M. F. and Clark, H. H. (1989). 'Understanding by addressees and overhearers', *Cognitive Psychology* 21(2): 211–232, https://doi.org/10.1016/0010-0285(89)90008-X.

Schouwstra, M. (2017). 'Temporal structure in emerging language: From natural data to silent gesture', *Cognitive Science* 41: 928–940, https://doi.org/10.1111/cogs.12441.

Schouwstra, M. and De Swart, H. (2014). 'The semantic origins of word order', *Cognition* 131(3): 431–436, https://doi.org/10.1016/j.cognition.2014.03.004.

Scott-Phillips, T. C. and Kirby, S. (2010). 'Language evolution in the laboratory', *Trends in Cognitive Sciences* 14(9): 411–417, https://doi.org/10.1016/j.tics.2010.06.006.

Scott-Phillips, T. C., Kirby, S., and Ritchie, G. R. S. (2009). 'Signalling signalhood and the emergence of communication', *Cognition* 113(2): 226–233, https://doi.org/10.1016/j.cognition.2009.08.009.

Segovia-Martín, J., Walker, B., Fay, N., and Tamariz, M. (2020). 'Network connectivity dynamics, cognitive biases, and the evolution of cultural diversity in round-robin interactive micro-societies', *Cognitive Science* 44(7): e12852, https://doi.org/10.1111/cogs.12852.

Selten, R. and Warglien, M. (2007). 'The emergence of simple languages in an experimental coordination game', *Proceedings of the National Academy of Sciences* 104(18): 7361–7366, https://doi.org/10.1073/pnas.0702077104.

Senghas, A. and Coppola, M. (2001). 'Children creating language: How Nicaraguan Sign Language acquired a spatial grammar', *Psychological Science* 12(4): 323–328, https://doi.org/10.1111/1467-9280.00359.

Senghas, A., Kita, S., and Özyürek, A. (2004). 'Children creating core properties of language: Evidence from an emerging Sign Language in Nicaragua', *Science* 305(5691): 1779–1782, https://doi.org/10.1126/science.1100199.

Shannon, C. E. and Weaver, W. (1949). *The Mathematical Theory of Communication*. Champaign, IL: University of Illinois Press.

Smith, A. D. M. and Höfler, S. H. (2015). 'The pivotal role of metaphor in the evolution of human language', in J. E. Díaz-Vera (ed.), *Metaphor and Metonymy across Time and Cultures*. Berlin: De Gruyter Mouton, https://doi.org/10.1515/9783110335453.123

Sperber, D. and Wilson, D. (1994). *Relevance: Communication and Cognition* (Reprint). Oxford: Blackwell Publishing Ltd.

Tamariz, M. (2017). 'Experimental studies on the cultural evolution of language', *Annual Review of Linguistics* 3(1): 389–407, https://doi.org/10.1146/annurev-linguistics-011516-033807.

Tamariz, M., Roberts, S. G., Martínez, J. I., and Santiago, J. (2018). 'The interactive origin of iconicity', *Cognitive Science* 42(1): 334–349, https://doi.org/10.1111/cogs.12497.

Trudgill, P. (2008). 'Linguistic and social typology', in J. K. Chambers, P. Trudgill, and N. Schilling-Estes (eds.), *The Handbook of Language Variation and Change*, 707–728. Oxford: Blackwell Publishing Ltd, https://doi.org/10.1002/9780470756591.ch27

Trudgill, P. (2009). 'Sociolinguistic typology and complexification', in G. Sampson, D. Gil, and P. Trudgill (eds.), *Language Complexity as an Evolving Variable*, 98–109. Oxford: Oxford University Press.

Verhoef, T., Roberts, S. G., and Dingemanse, M. (2015). 'Emergence of systematic iconicity: Transmission, interaction and analogy', *Proceedings of 37th Annual Meeting of the Cognitive Science Society*, 2481–2486.

Verhoef, T., Walker, E., and Marghetis, T. (2022). 'Interaction dynamics affect the emergence of compositional structure in cultural transmission of space-time mappings', *Proceedings of the 44th Annual Meeting of the Cognitive Science Society*, 2133–2139.

9780192886491_Book.indb 60 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:25

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Wilkes-Gibbs, D. and Clark, H. H. (1992). 'Coordinating beliefs in conversation', *Journal of Memory and Language* 31(2): 183–194, https://doi.org/10.1016/0749-596X(92)90010-U.

Winters, J., Kirby, S., and Smith, K. (2015). 'Languages adapt to their contextual niche', *Language and Cognition* 7(3): 415–449, https://doi.org/10.1017/langcog.2014.35.

Winters, J., Kirby, S., and Smith, K. (2018). 'Contextual predictability shapes signal autonomy', *Cognition* 176: 15–30, https://doi.org/10.1016/j.cognition.2018.03.002.

Wray, A. and Grace, G. W. (2007). 'The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form', *Lingua* 117(3): 543–578, https://doi.org/10.1016/j.lingua.2005.05.005.

Wu, S. and Keysar, B. (2007). 'The effect of information overlap on communication effectiveness', *Cognitive Science* 31(1): 169–181, https://doi.org/10.1080/03640210709336989.

Żywiczyński, P., Sibierska, M., Wacewicz, S., et al. (2021). 'Evolution of conventional communication: A cross-cultural study of pantomimic re-enactments of transitive events', *Language & Communication* 80: 191–203, https://doi.org/10.1016/j.langcom.2021.07.002.

9780192886491_Book.indb 61 /12_first_proofs/first_pr[illegible]/xml_for_typesetting 27-Mar-25 23:02:25

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

9780192886491_Book.indb 62 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:25

CHAPTER 4

VIRTUAL REALITY AS A TOOL TO STUDY LANGUAGE EVOLUTION

JONAS NÖLLE AND DAVID PEETERS

4.1 Introduction

If there is one thing we can safely assume when it comes to the evolution of language, it is that languages emerge and continue to evolve in physically, communicatively, and socially rich and dynamic situations. Indeed, humans naturally communicate using a variety of bodily channels (voice, hands, face) simultaneously, in close interaction with others and with aspects of the environment, while employing a range of complementary cognitive capacities (attention, memory, movement planning, executive control). How can one experimentally investigate such a complex phenomenon in the lab, ideally in settings that mimic the richness of everyday environments where people use language to communicate?

In this chapter, we present immersive Virtual Reality (VR) as a promising tool for studying language evolution in the lab, capturing its multifaceted nature. The chapter consists of three main sections. First, we introduce the reader to VR as a methodology, discussing its key concepts and affordances for language research in general. Second, we delve into the potential of VR for addressing theoretical research questions central to the study of language evolution, including a discussion of various experimental paradigms, multimodality, exploration of alternate realities, and animal communication. Third, we showcase a recent line of studies in this domain, illustrating how VR has been successfully employed to contrast and test different hypotheses in rich three-dimensional (3D) environments.

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

4.2 VR: Basic concepts and affordances for language research

Broadly, scientific studies using VR for stimulus display fall into two categories: those employing head-mounted displays (headsets) and those using a Cave Automatic Virtual Environment (CAVE). Head-mounted displays block the user's view of the real world and replace it with a 3D visual alternative through the headset, while a CAVE setup immerses participants in a stereoscopic virtual world via large digital screens or walls with projections experienced through 3D using shutter glasses (see Figure 4.1). In both cases, the virtual world adapts dynamically to participants' tracked behaviour, enabling them to experience the virtual world as real. 'Stereoscopic' refers to the digital environment being rendered from two slightly different perspectives, such as one for each eye in a headset, which simulates depth and creates a 'true 3D effect'. In contrast, monoscopic environments include 3D environments shown on a two-dimensional (2D) computer screen. In examples like 3D computer games, techniques such as shading and perspective lines create an illusion of depth, but they do not provide a true 3D experience, as is achieved with stereoscopic rendering.



FIGURE 4.1 (a) Example of a user wearing a VR headset to immerse himself into a virtual world. (b) Example of an experimental participant being immersed in a virtual forest in a CAVE, while her electrophysiological brain activity is continuously recorded via an electroencephalogram (EEG).

Source: (a) 'Exploring the Universe in Virtual Reality' by *NASA Goddard Photo and Video* is licensed under CC BY 2.0.

In essence, a convincing virtual experience should replace not only *visual* input but also align auditory and haptic feedback with the virtual environment. Headphones or audio speakers and haptic gloves or controllers can create a multisensory experience, while even smell and taste may be integrated. However, a purely virtual experience is challenging to achieve, as users' vestibular and proprioceptive sensations still derive from the real world (Skarbez et al. 2021). Thus, all present-day VR instances can actually be considered forms of *mixed reality*, where users experience a combination of virtual

9780192886491_Book.indb 64 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:25

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

and real-world sensations. Nevertheless, in this chapter, we will use the more popular term 'VR'. We note that VR differs from its technological counterpart, *augmented reality*, which uses the real world as its visible basis and adds digital elements through devices like transparent head-mounted displays or smartphones.

VR studies allow participants to perceive and interact naturally with the virtual world as if they were really there. For a successful study, participants must feel *present* in the virtual environment and find the illusion *plausible*, leading them to willingly *suspend disbelief*. A key distinction is between the objective notion of *immersion* and the subjective construct of *presence* (Cummings and Bailenson 2016). Immersion refers to the system's technical capability to deliver a convincing environment for participant interaction (Slater and Wilbur 1997; Sanchez-Vives and Slater 2005: 333). It is a system property that can be examined by inspecting technological aspects like field of view specs or quality of the tracking system. Conversely, presence refers to the subjective experience of feeling present in the virtual world and behaving accordingly (Heeter 1992; Slater 2009). Researchers often assess the extent of participants' spatial and social presence and perceived plausibility by using validated questionnaires post-study (e.g. Lessiter et al. 2001) or analysing how closely virtual behaviour resembles real-world behaviour (Slater 2004).

In a virtual world, the participant may encounter virtual equivalents of other living entities. Two often-confused concepts in this regard are *virtual agents* and *avatars*. Virtual agents refer to human-like digital characters controlled by a computer, showing pre-programmed behaviour. In contrast, avatars are human-like digital characters whose behaviour changes as a function of a human controller's concomitant actions. This distinction is reminiscent of the difference between player characters and non-player characters in video games. Recently, the terms *virtual humans* and *meta-humans* have gained popularity for describing both virtual agents and avatars.

So why would one use VR as a method to answer fundamental research questions, such as those in language evolution? Most prominently, VR has the potential to combine ecological validity and experimental control, offering the best of both worlds (see Figure 4.2). Participants can enter a virtual world that mimics the dynamics, interactivity, and multimodal, multisensory richness of the real world, creating an ecologically valid environment in the lab (e.g. Tromp et al. 2018; Peeters 2020; Nölle, Kirby, et al. 2020; Allritz et al. 2022). Treadmills may even allow exploration without real-world spatial limitations (Misersky et al. 2022). Simultaneously, researchers retain experimental control to manipulate one variable of theoretical interest between conditions and study the resulting effect on the data (kinematic, spoken, eye-tracking, electrophysiological) that are collected.

Additionally, the behaviour of pre-programmed virtual interlocutors is more consistent and replicable than the behaviour of human confederates, promoting reproducible research across participants and labs (Kuhlen and Brennan 2013; Pan and Hamilton 2018). Lastly, VR overcomes the artificial separation between stimulus and participant in 2D studies, allowing participants to quasi 'enter' the stimulus space (Peeters 2019).

9780192886491_Book.indb 65 /12_first_proofs/first_pr…/xml_for_typesetting 27-Mar-25 23:02:25

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

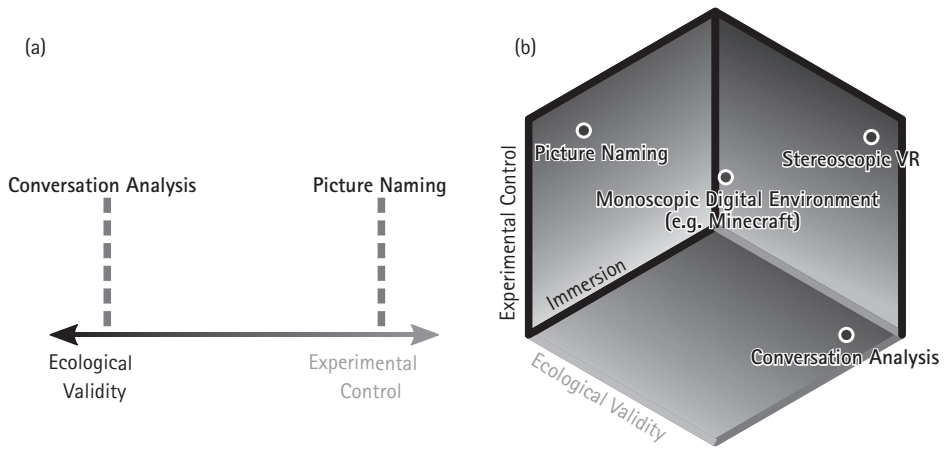


FIGURE 4.2 (a) Traditionally, ecological validity and experimental control have been conceptualized as opposite ends of a continuum. (b) In contrast, different experimental paradigms are represented as points in a three-dimensional coordinate system, where ecological validity (x-axis), experimental control (y-axis), and immersion (z-axis) form orthogonal aspects of experimental design.

Considering the opportunities VR brings, and in light of the technological advances made over the past decade, it is unsurprising that VR is quickly becoming a part of the experimental language researcher's methodological toolbox. VR has been employed to address various fundamental research questions in the language sciences, for instance, to examine how people align manual pointing gestures with speech (Chu and Hagoort 2014), how bilinguals codeswitch (Peeters and Dijkstra 2018), the role of embodiment in foreign language learning (Senkbeil et al. 2022), word prediction during listening (Huizeling et al. 2022), and the presence of motor activation during action verb processing (Zappa et al. 2019).

A decade of VR studies in the language sciences has methodologically demonstrated two key points. First, well-known psycholinguistic effects replicate in VR: research has shown syntactic priming between humans and virtual agents (Heyselaar et al. 2017), language switch costs for bilinguals (Peeters 2020), anticipatory eye movements during listening (Heyselaar et al. 2021), and a robust N400 effect following word–object mismatches (Tromp et al. 2018) in VR. Second, participants typically treat virtual agents similarly to how they would interact with actual humans (Gijssels et al. 2016; Heyselaar et al. 2017). These observations provide the necessary foundation for VR experiments to offer insights into real-world behaviour and processing.

Theoretically, VR has proven particularly valuable for studying multimodal and multichannel communication (e.g. Chu and Hagoort 2014; Hömke et al. 2018; Nota et al. 2022; Raghavan et al. 2023). Traditional studies often presented participants with only visual or auditory information, focusing on a single bodily channel, such as the voice or face. In contrast, VR studies naturally present multimodal information and

9780192886491_Book.indb 66 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:26

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

address communication as a multichannel phenomenon. This methodological richness forces researchers to consider the complexities of everyday communication and its environmental context. Consequently, VR is increasingly employed as a tool to study language evolution.

4.3 VR IN LANGUAGE EVOLUTION RESEARCH

Language evolution research has recently experienced an empirical turn, transitioning from speculative theorizing to an increasing focus on data-driven and experimental paradigms to study the origins and development of human language (see Fitch 2017; Tamariz 2017; Roberts 2018; Nölle, Hartmann, et al. 2020 for an overview). This shift recognizes the limitations of traditional comparative and phylogenetic approaches and embraces the use of broad datasets from various disciplines. In this context, VR has emerged as a powerful tool, enabling researchers to create controlled environments and implement novel communication systems for studying language evolution. The use of VR in language evolution research spans a wide range of topics.

In this section, we delve into the ways VR can enhance language evolution research, focusing on the benefits it brings to common experimental paradigms, such as artificial language learning and experimental semiotics tasks (see Chapters 1–3, 9, and 15, this volume). Additionally, we explore the impact of VR on the study of multimodality, animal communication, and the examination of language evolution within hyperrealities or possible worlds. These innovative approaches challenge traditional assumptions and open up new avenues for exploring questions in language evolution research. By integrating VR with cross-linguistic or cross-species fieldwork, as well as neural, genetic, and computational methods, researchers can make causal inferences, deepening our understanding of evolutionary mechanisms and driving further theoretical innovation in the field.

4.3.1 Common experimental paradigms

Language evolution research has made great strides using experimental paradigms such as artificial language learning, iterated learning, and experimental semiotics (Hudson Kam and Newport 2005; Galantucci 2005; Garrod et al. 2007; Kirby et al. 2008; see Galantucci et al. 2012; Tamariz 2017; Culbertson and Schuler 2019; Nölle and Galantucci 2022 for reviews). These approaches have allowed researchers to gain insights into the cognitive processes involved in language acquisition, language use in interaction, and language change over time. Despite their value in providing experimental control and simulating the natural dynamics of language emergence, these paradigms have sometimes been criticized for lacking ecological validity due to their use of highly abstract or artificial tasks (Tamariz 2017; Galantucci 2017).

9780192886491_Book.indb 67 /12_first_proofs/first_pr…/xml_for_typesetting 27-Mar-25 23:02:26

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

VR has the potential to significantly advance experimental language evolution research by offering a unique balance between experimental control, ecological validity, and immersion (see Figure 4.2). For example, VR can enhance symbol grounding tasks by providing a more realistic, multimodal environment while maintaining the same degree of control over the meaning space and referential environment traditionally found in experimental semiotics tasks (Galantucci et al. 2012; Nölle and Galantucci 2022).

In a VR setting, participants can engage in joint activities within a rich 3D environment that closely resembles real-world scenarios or even imaginary activities. For instance, symbol grounding tasks, like Galantucci's (2005) classic experiment that used a simple computer interface and a novel graphic communication medium, can be complemented by VR setups that provide a sweet-spot between realism/multimodal richness and control. Galantucci's setup arguably lacked ecological validity as it imposed some artificial restrictions on communication, such as participants not being co-present, thus unable to point at things to use as indices or icons, or to utilize alternative communication channels such as gestures. By adapting the game Minecraft for a dyadic computer-screen setup, Irvine and Roberts (2016) found that more immersive environments which provide these affordances can effectively remove the need for establishing symbols in favour of deictics. Extending this setup to an immersive, 3D VR environment would allow for even more naturalistic multimodal interactions.

VR settings can indeed incorporate a perfectly normal-looking environment containing everyday objects, offering high levels of immersion and ecological validity. At the same time, features of the referential environment can be controlled in the same way as in experimental semiotics tasks that usually rely on pictorial stimuli or computer interfaces. This enables researchers to investigate in more ecologically valid contexts how various features of the meaning space can lead to the emergence of structure in the signalling space, such as systematicity, underspecification, and category systems (cf. Winters et al. 2015; Carr et al. 2017; Nölle et al. 2018). VR thus provides the same amount of control but arguably sits higher on the immersion scale (z-axis in Figure 4.2b above) than the simple referential communication or coordination games often used to study language emergence via computer-screen setups. From a collaborative cooking or joint navigation task, essentially any conceivable activity can be implemented in VR.

A key advantage of VR compared to the use of a real physical space is reproducibility. In VR, objects can be instantiated in exact locations at exact times and their features can be controlled parametrically and continuously to create the exact referential environment the researcher is interested in. For instance, the number, size, shape, texture, or colour of objects appearing on a table could be varied randomly or in a discrete or continuous manner. A 3D engine, like Unity 3D, for instance, allows procedurally generating cups that interpolate between different sizes, colours, and textures (e.g. glossiness) that, at the beginning of each trial, can always appear in specific fixed positions on a table. One could then study how these features become encoded in a communication system that is grounded, used, transmitted, etc. This allows applying the traditional manipulations and structural organization of behavioural experiments—with blocks,

9780192886491_Book.indb 68 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:26

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

trials, conditions, randomization, or counterbalancing of items—directly to the multimodal world that participants are immersed in, and the entities they interact with.

4.3.2 Multimodality

The debate over the original modality of human language emergence, whether speech-first (Dunbar 1996; MacNeilage 2008) or gesture-first (Corballis 2003; Tomasello 2008; Arbib 2012; Levinson 2023), has given way to the understanding that language is inherently multimodal (Levinson and Holler 2014; Kendon 2017; Holler and Levinson 2019). Laboratory experiments have demonstrated the capability of gestures and non-linguistic vocalizations to ground novel signals and communicate various referents (Fay et al. 2014, 2022; Perlman et al. 2015; Ćwiek et al. 2021). However, different modalities have their own strengths and weaknesses in conveying meanings. Gestures, for instance, are often thought to be the most efficient means for symbol grounding, while vocalizations can convey emotions better than object-related information (Fay et al. 2014), and the iconic potential of facial expressions might be greater than usually assumed (Nölle et al. 2021).

VR experiments offer an ideal paradigmatic framework to systematically study the integration of modalities in real-time communication as well as the unique contributions and potential of different modalities in grounding novel communication systems. VR is well suited for studying non-verbal behaviour due to its embodied and immersive nature. Research has shown that people readily engage in non-verbal behaviour in VR settings (Maloney et al. 2020), and current technology allows for accurate rendering of real-life bodily movements, including posture, limb and finger movements, and even facial expressions produced under a VR headset onto avatars (Cha et al. 2020).

Using VR, researchers can also control and constrain available communication channels. For example, to adapt the classical 'silent gesture' paradigm (e.g. Goldin-Meadow et al. 2008; Motamedi et al. 2022) in VR, dyads can be placed in separate experimental booths with the voice channel disabled. For a vocalization-only condition, the headset microphone can be activated while disabling the rendering of the avatar's body or making it static. In a combined vocal-gesture condition (like in Fay et al. 2014), both channels can be enabled. Additionally, VR allows for the easy introduction of noisy conditions, such as making gestures systematically less visible by adding degrees of visual fog, altering the interpretability of vocalizations by manipulating the presence of environmental sounds, or applying algorithms to distort them (Rachman et al. 2018).

In sum, by employing VR in the study of multimodality in language evolution research, researchers can systematically explore the unique contributions of different modalities in communication and grounding novel signals, as well as investigate how these modalities interact and function as an integrated system. This approach can provide valuable insights into the complex dynamics of human communication and its evolution.

9780192886491_Book.indb 69 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:26

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

4.3.3 Strange and possible worlds

VR is often thought of as a simulation of reality. While the naturalism achievable in VR is indeed a strength, an even greater potential for research may lie in its ability to create situations and communication mediums beyond the affordances of reality as we know it. This extends from 'knocking out' modalities to designing entirely new communication systems.

For instance, VR facilitates the creation of a graphical communication system with strong affordances for iconicity going beyond the traditional 'Pictionary' experiments (Garrod et al. 2007). Participants can interact gesturally with three-dimensional canvases and manipulate how drawings appear to the interlocutor (see e.g. Google Tilt Brush). Researchers can control parameters like the rapidity of fading (Galantucci et al. 2010), the immediateness or displacement of communication (Nölle et al. 2018), or restrict iconicity using different drawing materials (Roberts et al. 2015). This opens the door for studying duality of patterning using tools like virtual slide whistles (Verhoef et al. 2014).

Beyond these novel communication mediums, VR allows for the exploration of hyper-realities or 'possible worlds'. In philosophy of mind and cognitive science, possible worlds are often used to study how we reason about counterfactuals or arrive at the semantics of natural language (Loux and Crisp 2017). Here, we mean the creation of alternate realities used to study the relationship between communication systems, innate cognitive bias, physical environments, and communicative niches, which taps directly into debates about the origins of linguistic structure (e.g. Christiansen and Chater 2008; Hurford 2011; Lupyan and Dale 2016; Kirby 2017; Berwick and Chomsky 2017).

Researchers can also examine how languages evolving in realities different from our own differ from natural languages, exploring the extent to which linguistic structures and universals derive from hard-wired rules and biases or domain-general cognitive abilities and interaction with the world. Hurford (2011) first proposed a potential experiment where objects frequently and rapidly morph into others—like a tree becoming a cow then a rock. This would challenge participants' linguistic concepts (such as what is typically referred to with nouns and verbs) and might lead to potentially inverted linguistic structures emphasizing constant actions over fleeting object identities when compared with the 'real world' (e.g. *spin pigging* vs. *pig spinning*; see Scott-Phillips et al. 2010).

A specific illustration of this approach can be found in the study of isomorphism in modifier ordering across languages, such as *these three green apples*. In both prenominal languages (like English) and post-nominal languages (like Thai), speakers tend to prefer demonstratives (e.g. *these*) to be further away from the noun than numerals, with adjectives being closest to the noun. This linguistic tendency seems to be explained by a cross-culturally robust cognitive bias for hierarchical structure (Culbertson and Adger 2014; Martin et al. 2019). However, the question remains where this bias comes from. Using VR, one could address this question empirically, and test the hypothesis that bias reflects the logical structure of the world. In an immersive

9780192886491_Book.indb 70 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:26

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

game, participants could interact with an 'alien world' that behaves slightly differently from our own to learn the names of alien objects and their properties in a post-nominal 'alien language'. Crucially, they could do so in one of two conditions, which differ in how inherent number (*one* or *many*) and distance (as a proxy for demonstratives like *this* and *that*) are to objects. For instance, if distance is more inherent, certain alien objects will always appear at a specific distance to the player, while their number varies (or vice versa). We train speakers of a pre-nominal language, like English, on post-nominal one-modifier constructions (e.g. X this, Y three) and can then test the word orders that participants spontaneously produce for two modifiers (Z three these vs. Z these three). If the preferred order reflects the more inherent feature, this proposed experiment would provide evidence for an empiricist (rather than nativist) account and suggest that this bias results from interaction with the physical world, constraining the noun phrase cross-linguistically.

In sum, by creating 'strange' worlds with varying properties, researchers can examine language evolution in alternate realities, testing the universality, flexibility, and robustness of linguistic structures and categories. Simulating 'alien worlds' with different properties and behaviours enables the exploration of unconventional questions and can challenge assumptions about natural language. How would languages conceptualize circular rather than linear time? Which frame of reference would interlocutors agree on to describe spatial locations if gravity was not reliable? How would languages encode colour or sizes of referents in a world where these features are not fixed? By testing how cognitive, biological, or environmental factors shape language evolution in alternate worlds and our own reality, we can directly address the problem of causality that remains one of the biggest challenges in the language sciences (e.g. Roberts 2018).

4.3.4 Animal communication

Understanding animal communication is essential for gaining insights into human language evolution, as it allows researchers to identify across species the biological, cognitive, and neural mechanisms that are necessary for complex communication (Fitch 2010, 2017). The use of virtual environments has proven to be a valuable tool for studying various animal species, including rodents and fish (Stowers et al. 2017).

A particularly promising paradigm in language evolution research involves using non-immersive touchscreen interfaces to create digital environments for non-human primates. This allows them to interact with, and solve, tasks in a virtual world presented on a 2D screen. Allritz et al. (2022) demonstrated that chimpanzees could be trained to navigate a large-scale, semi-realistic digital environment and successfully retrieve fruit items, which were then rewarded with corresponding real-world fruits (Figure 4.3a). This study showcases the potential virtual worlds have for examining spatial navigation in primates and its implications for the origins of spatial language in humans (Haun et al. 2006).

9780192886491_Book.indb 71 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:26

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Furthermore, virtual environments can be employed to investigate dyadic turn-taking and coordination in primates. Voinov et al. (2020) utilized a touchscreen interface in a dyadic task to study chimpanzees' ability to coordinate their behaviours for mutual benefit (Figure 4.3b). Their findings revealed that when communication was at risk of breaking down, chimpanzees would produce more spontaneous gestures to maintain coordination.

This highlights VR's potential to provide valuable insights into the evolutionary pathways of human language by examining the communication systems of other species. The versatility of game engines and 3D assets allows researchers to design similar scenarios for both humans and animals, facilitating the study of primate cognition, coordination, turn-taking, and communication in virtual worlds. This flexibility is particularly useful for primate studies, where providing large enclosures and variable environments as a basis for experimentation can be challenging and expensive. By using (2D) VR technology, researchers can create large-scale, controllable, and replicable environments that are crucial for studying specific phenomena and achieving generalizability.

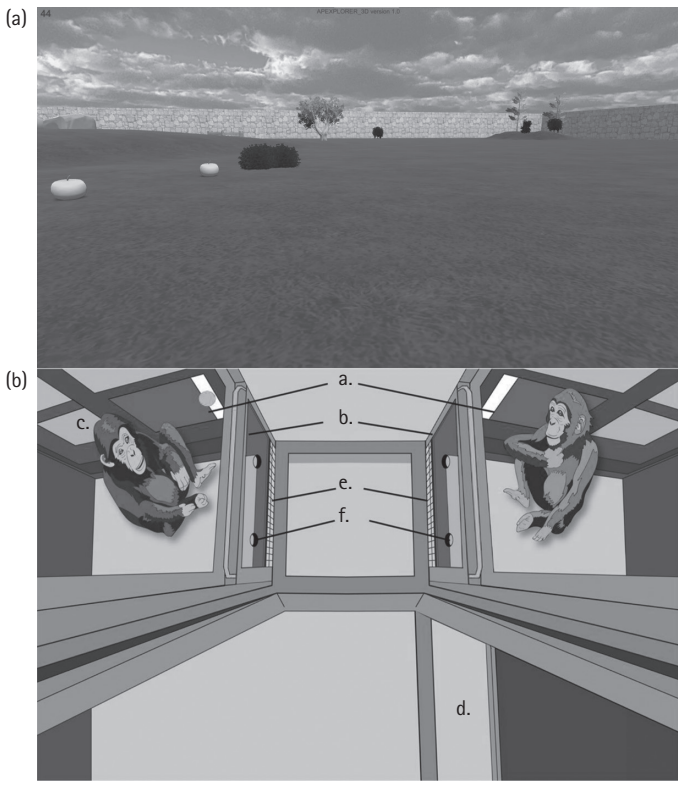


FIGURE 4.3 Chimpanzee touchscreen setups. (a) Allritz et al.'s (2022) navigation task, in which single apes can navigate a (monoscopic) virtual environment. (b) Voinov et al.'s dyadic setup, showing the touchscreens displaying stimuli (a.) and holes for gesturing (f.). For further details, see Voinov et al. (2020).

Source: Illustration by Marina Oshkina, reproduced with permission.

9780192886491_Book.indb 72 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:26

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

4.4 Case study: Using VR to study the evolution of spatial language

In this section, we will explore how VR can be employed to investigate concrete causal hypotheses in evolutionary linguistics by highlighting one recent series of VR studies in this domain. Recent years have seen a surge in research examining the adaptation of language to external variables, such as social, demographic, or environmental factors (Lupyan and Dale 2016). Examples include the proposed relationship between linguistic tone and humidity (Everett et al. 2015, 2016) and the negative correlation between morphological complexity and population size (Lupyan and Dale 2010). However, such cross-sectional studies often face criticism for their primarily correlational nature and lack of control for cultural evolution, which can confound data points (Roberts and Winters 2013). More recently, scholars have argued that robust, causal models of how external variables shape language require an integration of findings with methods that can directly test hypotheses, such as fieldwork, experiments, and computational models (Roberts 2018; Raviv et al. 2019; Nölle and Spranger 2022). For instance, Roberts (2018) suggests that a causal model of the relationship between humidity and tone necessitates converging evidence from laboratory phonetics, iterated learning experiments, corpus studies, and historical case studies.

One area where linguistic adaptation is fervently debated concerns the diversity in spatial language. Different speech communities exhibit preferences for various Frames of Reference (FoR) in linguistic and general cognitive tasks, such as an egocentric (*the ball is to the left of the car*) or an allocentric/geocentric strategy (*the ball is uphill of the car*). Recent research suggests that FoR choices could be predicted by language-external factors like topography, L2 contact, education, population density, and subsistence style (Bohnemeyer et al. 2015; Palmer et al. 2017; cf. Majid et al. 2004). However, one challenge is that these sociotopographic factors often intertwine and are difficult to disentangle (Nölle and Spranger 2022). For instance, in Marshallese and Dhivehi, topography's influence appears to be modulated by other factors. In both languages, the geocentric FoR seems motivated by the atoll topography, which strongly affords expressing directions such as *lagoonward* or *oceanside* compared to continental inland environments (Palmer et al. 2017). Yet, for Dhivehi, the geocentric FoR is more prevalent on fishing islands than non-fishing islands, and native Marshallese speakers in Springdale, Arkansas, have adopted egocentric terms from English, which they use in activities like driving (Schlossberg 2019).

A key challenge in attributing causality to such influences on spatial language lies in the difficulty of creating realistic, controlled settings for studying spatial referencing, a behaviour typically embedded in complex navigation activities. Spatial FoR experiments, however, often involve small-scale tasks, such as describing tabletop objects or photographs of spatial scenes, lacking immersion in large 3D environments where spatial referencing usually takes place. Lum and Schlossberg's (2014) Virtual Atoll

9780192886491_Book.indb 73 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:26

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Task attempted to address this issue with a (monoscopic) 3D navigation game playable on a laptop, prompting Dhivehi and Marshallese speakers to describe routes while navigating a digital atoll. However, the task was limited by its third-person perspective and the possible association of computers with left–right systems.

Nölle, Kirby, et al. (2020), therefore, designed a task to test the causal relationship between topography and spatial language in a laboratory setting, aiming to investigate the hypothesis that certain environmental features, such as the presence of landmarks enabling spatial descriptions, could drive systematic variation in FoR usage.

A two-player VR game called 'OrbHunt' was designed to study spatial language adaptation by simulating how players develop a spatial referencing strategy through dialogue and feedback over time. The game involved collecting yellow spheres, referred to as 'orbs', to score points. These orbs were only visible to the 'Seeker' player when they were in close proximity. The 'Director' player, on the other hand, could see the orbs from a distance and had to assist the Seeker in collecting them using spatial descriptions. After each round, players switched roles, enabling dyads to conventionalize strategies over time (Nölle 2021: ch. 4).

In Experiment 1, native English speakers wearing VR headsets were immersed in one of two large-scale environments, designed to motivate different spatial descriptions. The mountain slope had salient *up*, *down*, and *across* axes affording allocentric strategies, whereas the forest environment was dense and hard to orient, thus favouring egocentric solutions (Figure 4.4a). As predicted, Nölle, Kirby, et al. (2020) found that while English

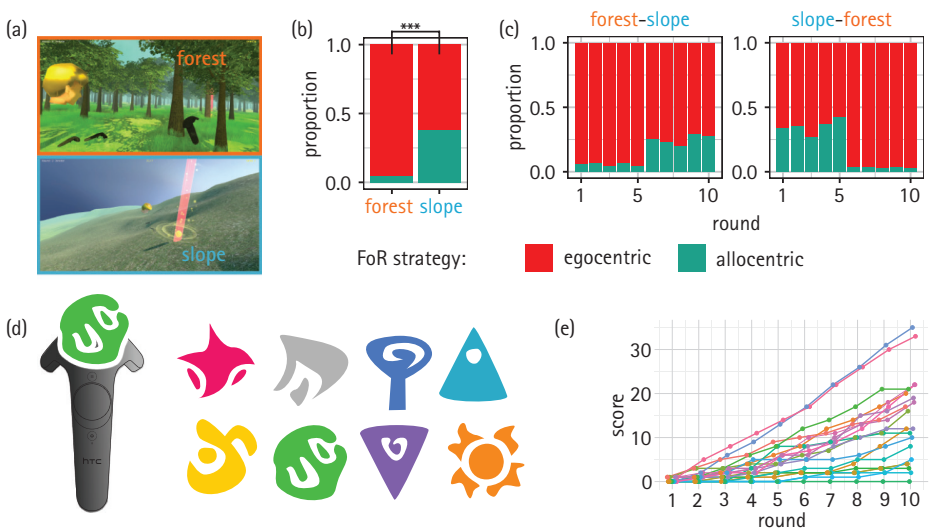


FIGURE 4.4 (a) The two virtual environments in the study by Nölle, Kirby et al. (2020). (b) Proportions of FoR strategy in both environmental conditions (Experiment 1) and (c) over time across both experimental blocks (Experiment 2). (d) The novel communication medium used in the experimental semiotics version of OrbHunt: the Director could select one of eight shapes on their controller and send them to appear in the Seeker's field of view. (e) Performance of dyads grounding a novel communication system.

9780192886491_Book.indb 74 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:27

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

speakers are generally regarded to favour egocentric solutions (Levinson 2003), which could be used in both environments, dyads on the slope used a significantly larger proportion of allocentric strategies (Figure 4.4b).

Experiment 2 tested how flexible these FoR strategies were when participants changed environment after five rounds. It was predicted that dyads switching from slope to forest should more likely abandon their initial geocentric strategy, while dyads switching from forest to slope could easily retain their egocentric strategy. However, the results showed that dyads readily adapted their strategies in both orders (Figure 4.4c). This suggests that speakers of languages that support several FoR readily and flexibly adapt to environmental cues, which could lead to the conventionalization of specific FoR strategies motivated by topography. Recently, a similar VR experiment by Tylén et al. (2022) that looked at transmissions of route descriptions in two different urban environments (a Manhattan-like grid vs. a more irregular 'Barcelona' condition) found that similar environmental influences emerged over time.

Nölle (2021) adapted the linguistic version of the OrbHunt game into an experimental semiotics task to explore the emergence and environmental adaptation of a completely novel spatial referencing system. In this version, dyads were restricted from using speech and instead, had to ground a new communication system in VR using visual signals that the Director could send to appear in the Seeker's visual field (Figure 4.4d). The experiment showed that although the task was more difficult, almost all dyads succeeded in grounding a spatial referencing system, with some dyads performing comparably to those playing the linguistic version of the game (Figure 4.4e). Additionally, head orientation and movement vectors of the players' avatars in relation to the signals in the game could be used to determine the underlying FoR for each signal with respect to the virtual environments (see Nölle 2021: ch. 5). The results revealed a strong egocentric bias, where many dyads routinized simple 'forward', 'backward', and 'turn left/right' signals that focused on steering the Seeker. Nölle (2021: 134ff.) interprets this as a natural consequence of the task emphasizing egocentric directions for navigation, similar to driving (Schlossberg 2019). To investigate what shapes diversity in FoR strategies, future experiments should include tasks on different scales that focus more on describing relative locations of objects, and also use participants who normally rely on allocentric systems to control for native language bias. Nevertheless, the experiment was a first demonstration of the emergence of a novel spatial communication system in the lab, and along with Experiments 1 and 2, adds further evidence to the adaptiveness of spatial language to both environments and specific human activities (Palmer et al. 2017; Tenbrink 2018).

The OrbHunt experiments collectively demonstrate the advantages of VR for studying spatial referencing, such as the visibility manipulations applied to the orbs and the director's avatar, and the large-scale nature of the environments that could not normally be achieved in the lab. Carefully controlled VR environments and various referential tasks could be used to test more explicit claims about how specific environments—such as slopes, atolls, rivers, and urban layouts—shape spatial language.

9780192886491_Book.indb 75 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:27

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

4.5 Conclusion

In conclusion, languages emerge and evolve within rich, dynamic, and interactive environments. Traditionally, it has been challenging to study language evolution with all its complexity and richness in an experimental lab setting while maintaining the experimental rigour required to collect reliable data. In this chapter, we have explored how VR can overcome such limitations by offering the potential to combine ecological validity with experimental control in reproducible and immersive research paradigms, addressing a wide variety of research questions pertinent to our understanding of language evolution.

The availability of commercial hardware, tutorials for setting up VR games using modern game engines (Murray 2017; Ferrone 2019), databases with standardized 3D objects (Peeters 2018; Tromp et al. 2018), and best practice guides for experimenters (Vasser and Aru 2020) have made this technology more accessible to researchers. While CAVE setups remain relatively expensive, VR consumer headsets have become affordable and offer high-quality 3D graphics, often in combination with built-in eye and head tracking technology. As we continue to push the boundaries of our understanding of language evolution, VR will undoubtedly play an increasingly vital role in shedding light on the complex interactions between linguistic, cognitive, and environmental factors that shape human communication. By leveraging the unique capabilities of VR, researchers can not only refine existing models and theories, but also explore previously inaccessible aspects of language evolution, driving the field towards new and exciting discoveries.

References

Allritz, M., Call, J., Schweller, K., et al. (2022). 'Chimpanzees (Pan troglodytes) navigate to find hidden fruit in a virtual environment', *Science Advances* 8(25): eabm4754, https://doi.org/10.1126/sciadv.abm4754.

Arbib, M. A. (2012). *How the Brain Got Language: The Mirror Systems Hypothesis*. Oxford: Oxford University Press.

Berwick, R. C. and Chomsky, N. (2017). 'Why only us: Recent questions and answers', *Journal of Neurolinguistics* 1–12, https://doi.org/10.1016/j.jneuroling.2016.12.002.

Bohnemeyer, J., Donelson, K. T., Moore, R. E., et al. (2015). 'The contact diffusion of linguistic practices', *Language Dynamics and Change* 5(2): 169–201.

Carr, J. W., Smith, K., Cornish, H., and Kirby, S. (2017). 'The cultural evolution of structured languages in an open-ended, continuous world', *Cognitive Science* 41(4): 892–923, https://doi.org/10.1111/cogs.12371.

Cha, H.-S., Choi, S.-J., and Im, C.-H. (2020). 'Real-time recognition of facial expressions using facial electromyograms recorded around the eyes for social virtual reality applications', *IEEE Access* 8: 62065–62075, https://doi.org/10.1109/ACCESS.2020.2983608.

9780192886491_Book.indb 76 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:27

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Christiansen, M. H. and Chater, N. (2008). 'Language as shaped by the brain', *Behavioral and Brain Sciences* 31(5): 489–509, https://doi.org/10.1017/S0140525X08004998.

Chu, M. and Hagoort, P. (2014). 'Synchronization of speech and gesture: Evidence for interaction in action', *Journal of Experimental Psychology: General* 143: 1726–1741, https://doi.org/10.1037/a0036281.

Corballis, M. C. (2003). 'From mouth to hand: Gesture, speech, and the evolution of right-handedness', *Behavioral and Brain Sciences* 26(2): 199–208, https://doi.org/10.1017/S0140525X03000062.

Culbertson, J. and Adger, D. (2014). 'Language learners privilege structured meaning over surface frequency', *Proceedings of the National Academy of Sciences* 111(16): 5842–5847, https://doi.org/10.1073/pnas.1320525111.

Culbertson, J. and Schuler, K. (2019). 'Artificial language learning in children', *Annual Review of Linguistics* 5(1): 353–373, https://doi.org/10.1146/annurev-linguistics-011718-012329.

Cummings, J. J. and Bailenson, J. N. (2016). 'How immersive is enough? A meta-analysis of the effect of immersive technology on user presence', *Media Psychology* 19(2): 272–309, https://doi.org/10.1080/15213269.2015.1015740.

Ćwiek, A., Fuchs, S., Draxler, C., et al. (2021). 'Novel vocalizations are understood across cultures', *Scientific Reports* 11(1): 10108, https://doi.org/10.1038/s41598-021-89445-4.

Dunbar, R. (1996). *Grooming, Gossip, and the Evolution of Language*. Cambridge, MA: Harvard University Press.

Everett, C., Blasi, D. E., and Roberts, S. G. (2015). 'Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots', *Proceedings of the National Academy of Sciences* 112(5): 1322–1327, https://doi.org/10.1073/pnas.1417413112.

Everett, C., Blasí, D. E., and Roberts, S. G. (2016). 'Language evolution and climate: The case of desiccation and tone', *Journal of Language Evolution* 1(1): 33–46, https://doi.org/10.1093/jole/lzv004.

Fay, N., Lister, C. J., Ellison, T. M., and Goldin-Meadow, S. (2014). 'Creating a communication system from scratch: Gesture beats vocalization hands down', *Frontiers in Psychology* 5: 1–12, https://doi.org/10.3389/fpsyg.2014.00354.

Fay, N., Walker, B., Ellison, T. M., et al. (2022). 'Gesture is the primary modality for language creation', *Proceedings of the Royal Society B: Biological Sciences* 289(1970): 20220066, https://doi.org/10.1098/rspb.2022.0066.

Ferrone, H. (2019). *Learning C# by Developing Games with Unity 2019: Code in C# and Build 3D Games with Unity*. Birmingham: Packt Publishing Ltd.

Fitch, W. T. (2010). *The Evolution of Language*. Cambridge: Cambridge University Press.

Fitch, W. T. (2017). 'Empirical approaches to the study of language evolution', *Psychonomic Bulletin & Review* 24(1): 3–33, https://doi.org/10.3758/s13423-017-1236-5.

Galantucci, B. (2005). 'An experimental study of the emergence of human communication systems', *Cognitive Science* 29(5): 737–767, https://doi.org/10.1207/s15516709cog0000_34.

Galantucci, B. (2017). 'Experimental semiotics', in Galantucci, B., *Oxford Research Encyclopedia of Linguistics*. Oxford: Oxford University Press, https://doi.org/10.1093/acrefore/9780199384655.013.210.

Galantucci, B., Kroos, C., and Rhodes, T. (2010). 'The effects of rapidity of fading on communication systems', *Interaction Studies* 11(1): 100–111, https://doi.org/10.1075/is.11.1.03gal.

Galantucci, B., Garrod, S., and Roberts, G. (2012). 'Experimental semiotics', *Language and Linguistics Compass* 6(8): 477–493, https://doi.org/10.1002/lnc3.351.

9780192886491_Book.indb 77 /12_first_proofs/first_pr[illegible]/xml_for_typesetting 27-Mar-25 23:02:27

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Garrod, S., Fay, N., Lee, J., et al. (2007). 'Foundations of representation: Where might graphical symbol systems come from?', *Cognitive Science* 31(6): 961–987, https://doi.org/10.1080/03640210701703659.

Gijssels, T., Casasanto, L. S., Jasmin, K., et al. (2016). 'Speech accommodation without priming: The case of pitch', *Discourse Processes* 53(4): 233–251, https://doi.org/10.1080/0163853X.2015.1023965.

Goldin-Meadow, S., So, W. C., Özyürek, A., and Mylander, C. (2008). 'The natural order of events: How speakers of different languages represent events nonverbally', *Proceedings of the National Academy of Sciences* 105(27): 9163–9168, https://doi.org/10.1073/pnas.0710060105.

Haun, D. B. M., Rapold, C. J., Call, J., et al. (2006). 'Cognitive cladistics and cultural override in Hominid spatial cognition', *Proceedings of the National Academy of Sciences* 103(46): 17568–17573, https://doi.org/10.1073/pnas.0607999103.

Heeter, C. (1992). 'Being there: The subjective experience of presence', *Presence: Teleoperators and Virtual Environments* 1(2): 262–271, https://doi.org/10.1162/pres.1992.1.2.262.

Heyselaar, E., Hagoort, P., and Segaert, K. (2017). 'In dialogue with an avatar, language behavior is identical to dialogue with a human partner', *Behavior Research Methods* 49(1): 46–60, https://doi.org/10.3758/s13428-015-0688-7.

Heyselaar, E., Peeters, D., and Hagoort, P. (2021). 'Do we predict upcoming speech content in naturalistic environments?', *Language, Cognition and Neuroscience* 36(4): 440–461, https://doi.org/10.1080/23273798.2020.1859568.

Holler, J. and Levinson, S. C. (2019). 'Multimodal language processing in human communication', *Trends in Cognitive Sciences* 23(8): 639–652, https://doi.org/10.1016/j.tics.2019.05.006.

Hömke, P., Holler, J., and Levinson, S. C. (2018). 'Eye blinks are perceived as communicative signals in human face-to-face interaction', *PLOS ONE* 13(12): e0208030, https://doi.org/10.1371/journal.pone.0208030.

Hudson Kam, C. L. and Newport, E. L. (2005). 'Regularizing unpredictable variation: The roles of adult and child learners in language formation and change', *Language Learning and Development* 1(2): 151–195, https://doi.org/10.1080/15475441.2005.9684215.

Huizeling, E., Peeters, D., and Hagoort, P. (2022). 'Prediction of upcoming speech under fluent and disfluent conditions: Eye tracking evidence from immersive virtual reality', *Language, Cognition and Neuroscience* 37(4): 481–508, doi:10.1080/23273798.2021.1994621.

Hurford, J. R. (2011). *The Origins of Grammar*. Oxford: Oxford University Press.

Kendon, A. (2017). 'Reflections on the "Gesture-First" hypothesis of language origins', *Psychonomic Bulletin & Review* 24(1): 163–170, https://doi.org/10.3758/s13423-016-1117-3.

Kirby, S. (2017). 'Culture and biology in the origins of linguistic structure', *Psychonomic Bulletin & Review* 24(1): 118–137, https://doi.org/10.3758/s13423-016-1166-7.

Kirby, S., Cornish, H., and Smith, K. (2008). 'Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language', *Proceedings of the National Academy of Sciences* 105(31): 10681–10686, https://doi.org/10.1073/pnas.0707835105.

Kuhlen, A. K. and Brennan, S. E. (2013). 'Language in dialogue: When confederates might be hazardous to your data', *Psychonomic Bulletin & Review* 20(1): 54–72, https://doi.org/10.3758/s13423-012-0341-8.

Lessiter, J., Freeman, J., Keogh, E., and Davidoff, J. (2001). 'A cross-media presence questionnaire: The ITC-sense of presence inventory', *Presence: Teleoperators and Virtual Environments* 10(3): 282–297, https://doi.org/10.1162/105474601300343612.

Levinson, S. C. (2003). *Space in Language and Cognition: Explorations in Cognitive Diversity*. Cambridge: Cambridge University Press.

9780192886491_Book.indb 78 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:27

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Levinson, S. C. (2023). 'Gesture, spatial cognition and the evolution of language', *Philosophical Transactions of the Royal Society B: Biological Sciences* 378(1875): 20210481, https://doi.org/10.1098/rstb.2021.0481.

Levinson, S. C. and Holler, J. (2014). 'The origin of human multi-modal communication', *Philosophical Transactions of the Royal Society B: Biological Sciences* 369(1651): 20130302, https://doi.org/10.1098/rstb.2013.0302.

Loux, M. J. and Crisp, T. M. (2017). *Metaphysics: A Contemporary Introduction*. New York: Routledge.

Lum, J. and Schlossberg, J. (2014). 'The Virtual Atoll Task: A spatial language elicitation tool', in M. Harvey and A. Antonia (eds.), *The 45th Australian Linguistic Society Conference Proceedings*, 82–103. Newcastle: NOVA Open Access Repository.

Lupyan, G. and Dale, R. (2010). 'Language structure is partly determined by social structure', *PLOS ONE* 5(1): e8559, https://doi.org/10.1371/journal.pone.0008559.

Lupyan, G. and Dale, R. (2016). 'Why are there different languages? The role of adaptation in linguistic diversity', *Trends in Cognitive Sciences* 20(9): 649–660.

MacNeilage, P. F. (2008). *The Origin of Speech*. Oxford: Oxford University Press.

Majid, A., Bowerman, M., Kita, S., et al. (2004). 'Can language restructure cognition? The case for space', *Trends in Cognitive Sciences* 8(3): 108–114.

Maloney, D., Freeman, G., and Wohn, D. Y. (2020). '"Talking without a voice": Understanding non-verbal communication in social virtual reality', *Proceedings of the ACM on Human-Computer Interaction* 4(CSCW2): 175:1–175:25, https://doi.org/10.1145/3415246.

Martin, A., Ratitamkul, T., Abels, K., et al. (2019). 'Cross-linguistic evidence for cognitive universals in the noun phrase', *Linguistics Vanguard* 5(1): 20180072, https://doi.org/10.1515/lingvan-2018-0072.

Misersky, J., Peeters, D., and Flecken, M. (2022). 'The potential of immersive virtual reality for the study of event perception', *Frontiers in Virtual Reality* 3, https://www.frontiersin.org/articles/10.3389/frvir.2022.697934.

Motamedi, Y., Wolters, L., Naegeli, D., et al. (2022). 'From improvisation to learning: How naturalness and systematicity shape language evolution', *Cognition* 228: 105206, https://doi.org/10.1016/j.cognition.2022.105206.

Murray, J. W. (2017). *Building Virtual Reality with Unity and Steam VR*. Boca Raton, FL: CRC Press.

Nölle, J. (2021). *How Language Adapts to the Environment: An Evolutionary, Experimental Approach* [University of Edinburgh], https://thesiscommons.org/bfuvr/.

Nölle, J. and Galantucci, B. (2022). 'Experimental semiotics: Past, present, and future', in A. M. García and A. Ibáñez (eds.), *The Routledge Handbook of Semiosis and the Brain*, 66–81. New York: Routledge.

Nölle, J. and Spranger, M. (2022). 'From the field into the lab: Causal approaches to the evolution of spatial language', *Linguistics Vanguard* 8(s1): 191–203, https://doi.org/10.1515/lingvan-2020-0007.

Nölle, J., Staib, M., Fusaroli, R., and Tylén, K. (2018). 'The emergence of systematicity: How environmental and communicative factors shape a novel communication system', *Cognition* 181: 93–104, https://doi.org/10.1016/j.cognition.2018.08.014.

Nölle, J., Hartmann, S., and Tinits, P. (2020). 'Language evolution research in the year 2020: A survey of new directions', *Language Dynamics and Change* 10(1): 3–26, https://doi.org/10.1163/22105832-bja10005.

9780192886491_Book.indb 79 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:27

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Nölle, J., Kirby, S., Culbertson, J., and Smith, K. (2020). 'Does environment shape spatial language? A virtual reality experiment', in A. Ravignani, C. Barbieri, M. Martins, et al. (eds.), *The Evolution of Language: Proceedings of the 13th International Conference (EvoLang13),* 321–323. Nijmegen: Max Planck Institute for Psycholinguistics.

Nölle, J., Chen, C., Hensel, L. B., et al. (2021). 'Facial expressions of emotion include iconic signals of rejection and acceptance', *Journal of Vision* 21(9): 2932, https://doi.org/10.1167/jov.21.9.2932.

Nota, N., Trujillo, J., and Holler, J. (2022). 'Conversational eyebrow frowns facilitate question identification: An online study using virtual avatars', *Cognitive Science* 47(12): e13392, https://doi.org/10.31234/osf.io/fcj8b.

Palmer, B., Lum, J., Schlossberg, J., and Gaby, A. (2017). 'How does the environment shape spatial language? Evidence for sociotopography', *Linguistic Typology* 21(3): 457–491, https://doi.org/10.1515/lingty-2017-0011.

Pan, X. and Hamilton, A. F. de C. (2018). 'Why and how to use virtual reality to study human social interaction: The challenges of exploring a new research landscape', *British Journal of Psychology* 109(3): 395–417, https://doi.org/10.1111/bjop.12290.

Peeters, D. (2018). 'A standardized set of 3-D objects for virtual reality research and applications'. *Behavior Research Methods* 50: 1047–1054.

Peeters, D. (2019). 'Virtual reality: A game-changing method for the language sciences', *Psychonomic Bulletin & Review* 26: 894–900, https://doi.org/10.3758/s13423-019-01571-3.

Peeters, D. (2020). 'Bilingual switching between languages and listeners: Insights from immersive virtual reality', *Cognition* 195: 104107, https://doi.org/10.1016/j.cognition.2019.104107.

Peeters, D. and Dijkstra, T. (2018). 'Sustained inhibition of the native language in bilingual language production: A virtual reality approach', *Bilingualism: Language and Cognition* 21(5): 1035–1061, https://doi.org/10.1017/S1366728917000396.

Perlman, M., Dale, R., and Lupyan, G. (2015). 'Iconicity can ground the creation of vocal symbols', *Royal Society Open Science* 2(8): 150152, https://doi.org/10.1098/rsos.150152.

Rachman, L., Liuni, M., Arias, P., et al. (2018). 'DAVID: An open-source platform for real-time transformation of infra-segmental emotional cues in running speech', *Behavior Research Methods* 50(1): 323–343.

Raghavan, R., Raviv, L., and Peeters, D. (2023). 'What's your point? Insights from virtual reality on the relation between intention and action in the production of pointing gestures', *Cognition* 240: 105581, https://doi.org/10.1016/j.cognition.2023.105581.

Raviv, L., Meyer, A., and Lev-Ari, S. (2019). 'Larger communities create more systematic languages', *Proceedings of the Royal Society B: Biological Sciences* 286(1907): 20191262, https://doi.org/10.1098/rspb.2019.1262.

Roberts, G., Lewandowski, J., and Galantucci, B. (2015). 'How communication changes when we cannot mime the world: Experimental evidence for the effect of iconicity on combinatoriality', *Cognition* 141: 52–66, https://doi.org/10.1016/j.cognition.2015.04.001.

Roberts, S. G. (2018). 'Robust, causal, and incremental approaches to investigating linguistic adaptation', *Frontiers in Psychology* 9: 166, https://doi.org/10.3389/fpsyg.2018.00166.

Roberts, S. G. and Irvine, E. (2016). 'Deictic tools can limit the emergence of referential symbol systems', in S. G. Roberts, C. Cuskley, L. McCrohon, et al. (eds.), *The Evolution of Language: Proceedings of the 11th International Conference*, http://evolang.org/neworleans/papers/99.html

Roberts, S. G. and Winters, J. (2013). 'Linguistic diversity and traffic accidents: Lessons from statistical studies of cultural traits', *PLoS One* 8(8): e70902.

9780192886491_Book.indb 80 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:27

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Sanchez-Vives, M. V. and Slater, M. (2005). 'From presence to consciousness through virtual reality', *Nature Reviews Neuroscience* 6(4): Article 4, https://doi.org/10.1038/nrn1651.

Schlossberg, J. (2019). *Atolls, Islands, and Endless Suburbia: Spatial Reference in Marshallese* [University of Newcastle], https://nova.newcastle.edu.au/vital/access/manager/Repository/uon:35604.

Scott-Phillips, T., Hurford, J. R., Roberts, G., and Roberts, S. (2010). 'Pragmatics not semantics as the basis for clause structure', in A. D. M. Smith, M. Schouwstra, de B. Boer, and K. Smith (eds.), *The Evolution of Language: Proceedings of the 8th International Conference (EVOLANG8)*, 491–492. Singapore: World Scientific, https://doi.org/10.1142/9789814295222_0108.

Senkbeil, K., Martin, G., and O'Rourke, B. (2022). 'SpEakWise VR: Exploring the use of social virtual reality in telecollaborative foreign language learning between learners of English and German', in B. Arnbjörnsdóttir, B. Bédi, L. Bradley, et al. (eds.), *Intelligent CALL, Granular Systems and Learner Data: Short Papers from EUROCALL 2022* (1st ed), 352–357. Research-publishing.net, https://doi.org/10.14705/rpnet.2022.61.1483

Skarbez, R., Smith, M., and Whitton, M. C. (2021). 'Revisiting Milgram and Kishino's reality-virtuality continuum', *Frontiers in Virtual Reality* 2: 1–8, https://www.frontiersin.org/articles/10.3389/frvir.2021.647997.

Slater, M. (2004). 'How colorful was your day? Why questionnaires cannot assess presence in virtual environments', *Presence: Teleoperators and Virtual Environments* 13(4): 484–493, https://doi.org/10.1162/1054746041944849.

Slater, M. (2009). 'Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments', *Philosophical Transactions of the Royal Society B: Biological Sciences* 364(1535): 3549–3557, https://doi.org/10.1098/rstb.2009.0138.

Slater, M. and Wilbur, S. (1997). 'A Framework for Immersive Virtual Environments (FIVE): Speculations on the role of presence in virtual environments', *Presence: Teleoperators and Virtual Environments* 6(6): 603–616, https://doi.org/10.1162/pres.1997.6.6.603.

Stowers, J. R., Hofbauer, M., Bastien, R., et al. (2017). 'Virtual reality for freely moving animals', *Nature Methods* 14(10): Article 10, https://doi.org/10.1038/nmeth.4399.

Tamariz, M. (2017). 'Experimental studies on the cultural evolution of language', *Annual Review of Linguistics* 3(1): 389–407, https://doi.org/10.1146/annurev-linguistics-011516-033807.

Tenbrink, T. (2018). 'Linguistic spatial reference systems across everyday domains: How people talk about space in sailing, dancing, and horse riding', *Cognitive Processing* 19: S31–S31.

Tomasello, M. (2008). *Origins of Human Communication*. Cambridge, MA: MIT Press.

Tromp, J., Peeters, D., Meyer, A. S., and Hagoort, P. (2018). 'The combined use of virtual reality and EEG to study language processing in naturalistic environments', *Behavior Research Methods* 50(2): 862–869, https://doi.org/10.3758/s13428-017-0911-9.

Tylén, K., Alinam, Z., and Vesper, C. (2022). 'Environmental and social factors in the emergence of spatial language', in A. Ravignani, R. Asano, D. Valente, et al. (eds.), *Proceedings of the Joint Conference on Language Evolution (JCoLE)*, https://doi.org/10.17617/2.3398549.

Vasser, M. and Aru, J. (2020). 'Guidelines for immersive virtual reality in psychological research', *Current Opinion in Psychology* 36: 71–76, https://doi.org/10.1016/j.copsyc.2020.04.010.

Verhoef, T., Kirby, S., and De Boer, B. (2014). 'Emergence of combinatorial structure and economy through iterated learning with continuous acoustic signals', *Journal of Phonetics* 43: 57–68.

9780192886491_Book.indb 81 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:27

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Voinov, P. V., Call, J., Knoblich, G., et al. (2020). 'Chimpanzee coordination and potential communication in a two-touchscreen turn-taking game', *Scientific Reports* 10(1): Article 1, https://doi.org/10.1038/s41598-020-60307-9.

Winters, J., Kirby, S., and Smith, K. (2015). 'Languages adapt to their contextual niche', *Language and Cognition* 7(3): 415–449, https://doi.org/10.1017/langcog.2014.35.

Zappa, A., Bolger, D., Pergandi, J. M., et al. (2019). 'Motor resonance during linguistic processing as shown by EEG in a naturalistic VR environment', *Brain and Cognition* 134: 44–57, https://doi.org/10.1016/j.bandc.2019.05.003.

9780192886491_Book.indb 82 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:28

CHAPTER 5

STUDYING CHILD-DIRECTED SPEECH

VERA KEMPE AND MARISA CASILLAS

To understand the mechanisms of language evolution, it is important to consider the nature of language input available to language learners. There is widespread acknowledgement that children's language input can differ considerably from the language used between adults. If language has evolved to be learnable (Christiansen and Chater 2008) then it is important to identify the proximate and ultimate mechanisms that render it so. One essential question, then, concerns the drivers of the modifications that characterize child-directed speech (CDS). To date, CDS has mainly been studied within the field of language development research, often with limited consideration of the proximate mechanisms that underpin its production. This chapter aims to describe the different questions that have been asked in the study of CDS and the methodological approaches taken to answer them. We will try to illuminate what the available evidence tells us about the mechanisms and role of CDS in underpinning the transmission of language as a cultural trait. The chapter concludes by pointing to future research that can advance understanding of how CDS fits into the broader concept of multimodal child-directed communication and how language transmission is embedded in the diversity of children's wider linguistic, physical, and social environment.

To begin, some terminological clarification is in order. Although the research literature features different labels for CDS, e.g. 'motherese', 'parentese', 'babytalk', 'caregiver speech', 'child-directed communication', or 'infant-directed speech', each of them emphasizing somewhat different perspectives on the phenomenon, we will use CDS as a historically apt shorthand for the multitude of ways in which children at different ages and in different circumstances may be addressed by primary and alloparental caregivers. Despite the use of the term CDS, we acknowledge that child-directed communication goes beyond spoken language and includes child-directed input in sign languages (e.g. Holzrichter and Meier 2000) as well as non-verbal behaviours that accompany verbal communication like gesture, movement, and tactile information (e.g. Brand and Shallcross 2008; Abu-Zhaya et al. 2017).

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

The study of CDS started in the 1960s, when the field of child language development research was dominated by a theoretical account that viewed language input as merely a trigger of maturational processes driven by an innate, biologically determined capacity (Chomsky 1957). Yet, already in 1964, Ferguson noted cross-linguistic parallels in the existence of phonological simplifications of some words caregivers use when addressing infants and small children. He suggested that these words serve two functions: to evoke nurturant behaviours and to aid language learning (Ferguson 1964). Because of the characteristic differences to adult-directed speech (ADS), the term CDS typically came to denote not just the language that is directed at children but a distinct way of speaking, akin to a speech register (but see Saxton 2008 and Weinstein and Baldwin 2022 for a critical analysis of the applicability of this sociolinguistic concept to CDS). The conceptualization of CDS as a distinct form of speech sparked a prolific line of research into the features of CDS, the factors that influence the propensity of caregivers to modify their language when addressing small children and the effects of these modifications on language development.

5.1 Comparing features of CDS and ADS

One prominent strand of research on CDS has aimed to characterize what distinguishes it from ADS. Much of this work does so by adopting the notion that CDS is a special form of speech rather than just a label for the part of child-surrounding language that happens to be directed at the child. Descriptions of the modifications taking place in CDS have been documented even though not every linguistic community has a distinct register for talking to children, a point we will return to below. Prior work in this research strand has typically obtained aggregate measures of one or several features, e.g. pitch, pitch range, articulation rate, vowel length, voice onset time of stop consonants, formant frequencies of vowels, mean length of utterances or type/token ratios of lexical items, to name a few, and compared them between CDS and ADS. The observed differences tend to form the basis for theory building about putative functions of CDS. In this section, we discuss research that focuses on comparing CDS and ADS but does not test the learning outcomes associated with the observed differences. Conclusions about the function of CDS are drawn based on theoretical assumptions about the potential effects of CDS features, e.g. by stipulating that those features that differentiate CDS from ADS can be assumed to benefit parent–child interaction, communication, and language development. In section 5.3 we discuss research that explicitly examines whether learning gains can indeed be attributed to the specific characteristics of CDS.

5.1.1 Comparing prosody and phonology

To date, most studies that compare CDS and ADS have focused on prosodic and phonological features. These studies typically analyse naturalistic recordings of CDS,

9780192886491_Book.indb 84 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:28

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

either obtained in the home environment or in the laboratory. The ADS samples used for comparison tend to be recorded during conversations between caregivers and experimenters or, less frequently, other family members thereby mainly employing within-participant designs for the CDS–ADS comparison. Recordings typically span a timeframe of minutes rather than hours and are subjected to acoustic analyses using software such as PRAAT (Boersma and Wenink 2005), to measure parameters like mean, minimum, and maximum fundamental frequency (F0), vowel duration, or pause duration. For comparisons to be valid, many studies adopt the approach of selecting CDS and ADS samples that are similar in context, content, and structure. To achieve this, caregivers may be asked to talk about the same child-centred actions or objects when addressing their child and an adult. One of the first studies of this type compared prosodic features of CDS and ADS directed at the other parent and the experimenter in mothers and fathers speaking Japanese, German, Italian, French, British or US English, and documented characteristic features like elevated pitch, exaggerated pitch range, and slower speech rate in CDS (Fernald et al. 1989). Pitch range was found to be wider in maternal compared to paternal CDS, and all prosodic features of CDS were most exaggerated in US English. This study seeded a prolific line of research into comparing CDS with ADS prosody across a range of different languages (Cox et al. 2022).

The study of phonological features of CDS was inspired by Lindblom's (1990) notion of hyperarticulation that is deployed to accommodate the needs of listeners. Kuhl et al. (1997) were the first to demonstrate increased vowel space, encompassed by the triangle of the 'corner' vowels /i/, /a/, and /u/ in first and second formant space, in English, Russian, and Swedish CDS. Their findings cemented the idea that CDS serves to support the learning of phonological categories based on statistical distributions of segmental cues, an idea that became known as the *hyperspeech* hypothesis of CDS. However, attempts to generalize the *hyperspeech* hypothesis to other phonemes and segmental contrasts have yielded inconsistent results. Some CDS–ADS comparisons of phonemically contrastive vowels situated inside the vowel triangle provided evidence for hypoarticulation and for increased variability in the instantiations of individual vowels (e.g. Cristia and Seidl 2014). Comparisons of voice onset times, by and large, failed to show evidence for enhancement of the contrast between voiced and voiceless stop consonants in CDS (Baran et al. 1977; Sundberg and Lacerda 1999; Englund 2005; McMurray et al. 2013), and research on lexical tones in languages like Mandarin or Cantonese failed to produce evidence for enhanced tonal contrasts (Wong and Ng 2018). More broadly, perceived speech clarity does not appear to reliably distinguish child-directed from overhearable speech in naturalistic US English (Foushee et al. 2023). This lack of evidence for a generally enhanced discriminability of phonemes suggests that hyperarticulation is probably a heavily contextualized phenomenon. Its variability across phonemes and languages may be explained by trade-offs between prosodic and phonological features of CDS. For example, variability in vowel instantiations and reduced tonal contrasts may be a consequence of the exaggerated pitch range of CDS (Cristia and Seidl 2014; Wong and Ng 2018), and the general lengthening of voice onset times in CDS may be a by-product of its slower articulation rate (McMurray et al. 2013).

9780192886491_Book.indb 85 /12_first_proofs/first_pr…/xml_for_typesetting 27-Mar-25 23:02:28

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Inconsistency in findings with respect to hyperarticulation in CDS may be due to a range of variables that are likely to moderate the distribution of phonological cues in CDS. Some of these potential moderators are the age of the child, the timing of emergence of certain words in the child's vocabulary, the lexical context of phonemes, the position of target words within utterances, or how liable certain phonemes are to be modulated by prosody (e.g. sonorants or lexical tones are more likely to be affected by pitch modulations than obstruents). Researchers have tried to control some of these contextual variables through elicited CDS where caregivers are invited to talk about preselected toys or picture books in an attempt to control the lexical context, often at the expense of ecological validity. One quite labour-intensive solution to this problem involves scanning large corpora of CDS and ADS recorded in naturalistic settings and selecting a subset of utterances that are matched on a range of criteria. Using this method, Englund (2005) revealed that VOT differentiation may be increased in CDS compared to ADS for some contrastive pairs of voiced and unvoiced consonants but not others.

5.1.2 Comparing morphosyntax, vocabulary, and pragmatics

The earliest studies of CDS typically relied on informants to document the lexical, syntactic, and discourse differences between CDS and ADS, mainly by cataloguing the specific words that were deemed to be unique to this register (e.g. reduplications, onomatopoeia, diminutives, or hypocoristics; see Ferguson 1964; Oswalt 1976). Subsequent studies compared corpora of naturalistic CDS and ADS in terms of frequency of certain words or constructions thought to make the task of extracting statistical regularities easier for the child, e.g. antonyms (Jones 2007), diminutives (Kempe et al. 2001), or specific parts of speech (Adi-Bensaid et al. 2015).

With the emergence of digital storage solutions for standardized transcripts of speech corpora through the creation of the CHILDES database (MacWhinney and Snow 1985) researchers were able to share naturalistic CDS corpora more widely. At the time of writing, the CHILDES database contains approximately 436 corpora comprising over seventy million words uttered by over 16,000 monolingual and multilingual children in forty-eight languages. Because corpora that document naturalistic interactions between children and caregivers (as opposed to elicited narratives) contain roughly twice as much speech by caregivers than speech by children, CHILDES has become a formidable repository of CDS. However, the nature of CHILDES as a repository of caregiver–child interactions does not readily provide for an ADS comparison. As a result, comparable ADS corpora need to be obtained from other recorded adult–adult conversations, e.g. from conversations between college students or from written sources. Only when comparable ADS corpora have been sourced can comparisons to CDS from the existing corpora be conducted.

Comparisons to ADS to date have shown that CDS directed to very young children tends to differ in some general measures such as mean length of utterance, often taken

9780192886491_Book.indb 86 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:28

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

to indicate syntactic complexity, type–token ratios of individual words as a measure of lexical complexity or frequency of certain speech acts as manifestations of caregiver responsiveness, sometimes all in the same study (e.g. Hayes and Ahrends 1988). Studies comparing more specific features between CDS and ADS have shown that CDS is generally characterized by simplified structural features, for example, higher frequency of diminutives (Kempe et al. 2001) or being more associative and repetitive but less contextually diverse (Hills 2013; Lester et al. 2022). However, to ascertain what role features that distinguish CDS from ADS may play in language development, it is necessary to examine their effects on learning gains, as we will discuss in section 5.3. A deeper understanding of the function of CDS also benefits from comparing it to ADS sampled from a greater variety of contexts and situations (Weinstein and Baldwin 2022), and to other forms of speech directed at different addressees like foreigner-directed speech (e.g. Piazza et al. 2022), pet-directed speech (e.g. Burnham et al. 2002), or elderspeak (e.g. Kemper and Harden 1999).

In contrast to the rich literature that compares prosodic and phonological features between CDS and ADS, studies devoted to a direct comparison of morphosyntactic, lexico-semantic, pragmatic, and discourse features are few and far between. We suggest several reasons for this discrepancy in the evidence base for CDS–ADS comparisons across linguistic domains: first, as mentioned above, the CHILDES database serves as a rich repository of speech by children and caregivers but has not been set up to document speech between adults. Second, studies of prosodic and phonological features can get by with relatively short recordings, often elicited in the lab, to obtain enough utterance and phoneme tokens to enable reliable comparisons of relevant acoustic parameters. Studies of referential and structural aspects of CDS, on the other hand, require recordings that span multiple hours to extract sufficient distributional information for a reliable comparison, which renders such studies much more cost- and labour-intensive and makes it more difficult to employ within-participant study designs.

Finally, the kinds of situational contexts typically chosen to obtain CDS recordings (e.g. playtime, mealtime, grooming, and bedtime routines) differ markedly from the situational contexts in which adults tend to interact with each other. In most studies, ADS corpora tend to differ in context and content and therefore introduce a range of situational and contextual confounds into the comparison (Weinstein and Baldwin 2022). Even if children and adults are placed in similar situations, speech to young children tends to focus on the here and now, while adults often talk about topics displaced in space and time (Foushee et al. 2023). This makes it difficult to tease apart whether potential differences in structural features and in referential content are due to the addressee or due to the topic and the situational context, suggesting that CDS is to a large extent situationally and culturally constrained. One solution to this problem of comparison, adopted from the study of prosodic and phonological properties (Liu et al. 2009), involves letting parents observe video recordings of other parents interacting with their children and discussing this child-centred content with another adult (e.g. Adi-Bensaid et al. 2015), thereby equating the content of both registers to some degree. Another solution is to elicit storytelling from adults directed at different addressees, a

9780192886491_Book.indb 87 /12_first_proofs/first_pr/xml_for_typesetting 27-Mar-25 23:02:28

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

method that has shown greater redundancy in speech addressed to linguistic novices, be it children or adult language learners (Tal et al. 2023).

5.2 Studying variation in CDS

Despite the initial suggestions that CDS might be a human universal (Ferguson 1964, 1977; Fernald 1992), there appears to be considerable variation in its prevalence and manifestation across different societies, languages, sociodemographic strata, or in response to the needs of non-typically developing children. Studying this variation can help to pinpoint conditions under which CDS can emerge as a culturally transmitted trait.

5.2.1 Cross-cultural variation in CDS

Ethnographic[1] accounts of early language learning and language socialization were the first to reveal that CDS as a specialized register is not apparent in all communities (Ochs 1982; Heath 1983; Schieffelin 1985), challenging the notion of universality of CDS. To gain a fuller picture, recent research has focused on documenting the prevalence of CDS across different societies and cultural groups more systematically. Cross-cultural comparison studies typically measure either the amount and duration of time intervals during which children are being addressed directly or the number of word tokens, word types, or whole utterances directed at children. Prevalence estimates require representative sampling of the language surrounding the child. One way of achieving this is through systematic observations and coding of presence or absence of CDS within pre-specified regular time intervals (e.g. Konner et al. 1977) or through time-allocation methods where all activities of target individuals, including speech production, are systematically recorded throughout the day (e.g. Cristia et al. 2019), though this method may underestimate CDS, as in Scaff et al. (2022).

Representative estimates of CDS prevalence across communities can also be obtained from video recordings of verbal interactions between caregivers and children taken in pre-specified situations such as playtime, mealtime, or other routines (e.g. Shneidman and Goldin-Meadow 2012; Vogt et al. 2015; Tamis-LeMonda et al. 2019). More recently, day-long (aka long-form) audio recordings are being used to capture larger samples of ambient language without having to take *a priori* decisions about which situations or

[1] Ethnographic research is most closely associated with the field of anthropology. It typically involves a systematic qualitative description of the study phenomenon, situated in its lived context. It aims to consider the perspectives of the individuals who relate closely to the phenomenon as well as a full array of salient, historical, institutional, and other factors affecting how the phenomenon is observed and experienced.

9780192886491_Book.indb 88 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:28

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

time intervals to observe or record (see Pisani et al. 2021 for a detailed review). This approach is often supported by the Language ENvironment Analysis system (LENA™) which allows researchers to obtain audio recordings with a device that has sufficient battery capacity for sixteen hours of continuous recording and can be fastened to the child's chest in a piece of clothing. The LENA™ proprietary software provides automated estimates of word counts in nearby adult speech, estimated counts of vocalizations produced by the target child, and estimated counts of speaker switches between the target child and a nearby adult (i.e. instances of turn-taking). Validation studies of the algorithms for deriving these automated measures, which have been trained on US English, show that the LENA™ system performs fairly well across many of the languages for which it has been used in the past. However, caution should be taken in using it with languages for which there is no existing validation. A more recent open-source Voice Type Classifier (Lavechin et al. 2020; Räsänen et al. 2021) provides many of the same measures but is trained on more linguistically diverse data though, again, validation is key to reliably interpreting automated results of this sort. In general, automated approaches can provide data on a previously unreachable scale and, when carefully employed, bring about groundbreaking insights. For example, recent work using over 40,000 hours of LENA™ recordings from 1,001 children across six continents revealed that, while children's speech-like vocalization rate was predicted by input rate and normative development, rates of nearby adult speech (likely often CDS) were not predicted by any of the usual suspected factors like child age, gender, SES, multilingualism, or language (Bergelson et al. 2023).

To be clear, automated analyses are not the only approach available for day-long recordings—another option is to subject recordings to acoustic, linguistic, or image-based (for video recordings) analyses (e.g. Ko et al. 2016; Montag 2020; Casey et al. 2022). Still, sometimes arduous, manual annotations of day-long data that go beyond what can currently be automatized are essential for revealing how CDS and other input sources differ between the lab and real home settings (Bergelson et al. 2019), but there is rapid progress in automated solutions for transcribing and converting audio into standardized analysis formats (Liu et al. 2023).

Because of its unobtrusiveness, the day-long audio-recording methodology is increasingly being adapted to the study of CDS beyond North America and western Europe (Henrich et al. 2010; Singh et al. 2023). For example, day-long recordings have been used to study prevalence of CDS in rural societies, including populations of subsistence farmers, farmer-foragers, and populations transitioning from these occupations into a market integrated with the global economy. Some examples include studies of children learning Tsimane' (isolate) in the lowlands of Bolivia (Scaff et al. 2022), Quechua (Quechua)-Spanish (Indo-European) bilinguals in the highlands of Bolivia (Cychosz et al. 2021), Yélî Dnye (isolate) learners on Rossel Island, Papua New Guinea (Casillas et al. 2021), and children learning Tseltal (Mayan) in southern Mexico (Casillas et al. 2020). Comparisons of input estimates in comparably constructed cross-cultural corpora (Bunce et al. 2024) have so far revealed that, at the day-long scale, rates of speech directed to children from adults are surprisingly similar across diverse societies,

9780192886491_Book.indb 89 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:28

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

however, there are clear differences in the availability of overhearable speech and CDS from other children. That said, input rates are sensitive to how researchers sample the input and what is included as directed input, with dramatic differences in estimates from one technique to another (Cristia 2019; Scaff et al. 2022).

Without a doubt, children have vastly different experiences learning language across diverse human societies—the ethnographic studies cited at the start of this section describe several such cases. Whether and how these differences may be picked up at the day-long scale and by automated measures is a challenge for future work. Urban and rural communities differ on a number of dimensions (birth rates, parenting strategies, presence of other children, expectations about children's capabilities and roles, education and literacy, type and division of adult labour) that are difficult to disentangle with traditional psychological methods (Cristia 2019). Work that adequately addresses cross-cultural differences will, in the future, likely need to include significant ethnographic and/or qualitative elements. So far, the number of non-Euro-American societies studied is still too small to draw definitive conclusions about which factors moderate the prevalence or features of CDS.

Cross-cultural and cross-linguistic differences in CDS do not preclude the existence of some similarities. Indeed, the manifestation of CDS has sometimes been found to have similar attributes across unrelated linguistic communities. For example, prosodic contours of US English CDS expressing different intentions like prohibition, approval, attention, and comfort are recognizable in South American Shuar hunter-horticulturalists (Bryant and Barrett 2007). A careful study of K'iche' (Mayan, Guatemala) CDS revealed a distinct register for addressing children that had some features in common with US English CDS (e.g. use of diminutives), but not others (high or variable pitch; Pye 1986). A recent meta-analysis of prosodic and phonological features of CDS from eighty-eight studies confirmed increased pitch, pitch range, vowel duration, vowel space, and reduced speech rate in the CDS in over thirty geographically diverse language communities (Cox et al. 2022). Yet, there was also considerable variation which cannot be explained entirely by methodological differences like whether CDS was observed in naturalistic vs. laboratory settings or whether it was read vs. produced spontaneously. This apparent variation around a general trend calls for further study of what drives cross-cultural differences in prevalence and manifestations of CDS.

5.2.2 Studying variation within linguistic communities

In a landmark study, Hart and Risley (1992, 2003) combined data from ethnographic observations and transcribed video recordings of CDS from forty-two US primary caregivers, estimating that high-SES children heard, cumulatively, around thirty million more word tokens than their low-SES peers over the first few years of life. This estimate prompted widespread concerns about linguistic deprivation associated with poverty and its potential long-term consequences for language development and academic

9780192886491_Book.indb 90 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:28

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

achievement. Yet, when trying to replicate this finding in several low- and middle-SES communities, Sperry et al. (2019) found considerable variation in language input to children, especially when taking into account overheard speech. A recent meta-analysis concluded that while published studies collectively support the idea that CDS quantity varies by SES, the differences are of a much smaller magnitude than estimated by Hart and Risley (1992, 2003) and vanish when all child-surrounding speech is taken into account (Dailey and Bergelson 2022). That finding is consistent with the aforementioned much larger study of over 40,000 hours of children's home language experiences, which found no effect of SES on child-surrounding adult speech quantity (Bergelson et al. 2023). It is also important to put the observed group differences in perspective: differences in CDS prevalence between urban and rural communities across cultures tend to exceed the ones observed between high and low socioeconomic strata within Western industrialized societies (Cristia 2022).

As we have seen, for many children, CDS may not be the primary source of language input. Yet, what children learn from surrounding speech not directly addressed to them is far from clear (Akhtar 2005; Weisleder and Fernald 2013; Foushee et al. 2023) and may depend on factors such as child age, what is being learned (e.g. words vs. facts. vs. prosodic patterns) and what attentional stance is encouraged in each culture and community. For example, while joint attention during dyadic interaction is encouraged in middle- and high-SES Western communities, other communities may expect children to learn from observing how communication is conducted by others (Rogoff et al. 1993; Lancy 2015). Future research should prioritize the study of the different paths to language that result from variation in how child-surrounding language is distributed across speakers and communicative contexts.

More recent approaches are moving away from the focus on CDS 'quantity' as a predictor of language development by emphasizing the importance of 'quality', particularly the content of caregivers' contingent responses to scaffold child language beyond the current point in development (Rowe 2012; Masek et al. 2021). It can be argued, however, that the quantity-vs.-quality debate distracts from the real goal of research in this domain: to gain a contextualized perspective on the role of children's language input (Casillas 2023). Caregivers talk differently from family to family, not just due to differences associated with SES (Hoff-Ginsberg 1991), but also differences in linguistic and cultural context, as well as in immediate circumstance. All these factors shape the ways in which children experience language in their social worlds.

In addition to variation in CDS across social groups, there is also variation associated with different types of clinical delays or disorders relating to communication and language—sometimes arising spontaneously and sometimes arising from clinical advice that parents are given. At present, most research of this type concerns CDS directed to children diagnosed with, or at high assessed risk of, Autism Spectrum Disorder (ASD, primarily verbal autistic children), Down syndrome, and other learning impairments, as well as children with hearing impairments (especially following receipt of hearing aids or cochlear implants). Surprisingly few studies have been conducted on speech directed to blind children (Campbell and Bergelson 2022; Perez-Pereira and Resches

9780192886491_Book.indb 91 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:28

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

2008) or children who stutter (Zebrowski et al. 1996). In this work, researchers typically compare CDS addressed to children diagnosed with one of these clinical conditions to CDS addressed to age- and/or language-matched peers not diagnosed with any clinical conditions relating to language or communication.

These studies, while often carried out in clinical settings with the aim to optimize language provision, are also instructive about deficit views of development—they inform us about how parental and clinical assumptions about children's capacities can shape children's language input, including CDS, and give us insight into the ways in which caregivers are able to accommodate a wider range of children's communicative behaviours. For example, when addressing verbal autistic children, caregivers have been shown to display increased responsiveness compared to interactions with non-autistic children. They have also been shown to produce more information-salient structural features such as yes/no questions, lower mean lengths of utterances, or more concrete nouns (Curcio and Paccia 1987; Konstantareas et al. 1988; Walton and Ingersoll 2015; Tubul-Lavy et al. 2020; Bentenuto et al. 2021; Onnis et al. 2021), which are indicative of their accommodation to the communicative needs of different children. Children with Down syndrome and other learning disabilities, on the other hand, while being exposed to more exaggerated prosody (Fidler 2003), may encounter less child-appropriate transparent word labelling, less decontextualized language, and fewer questions—features that may reflect an underestimation of these children's cognitive capacities (e.g. de Falco et al. 2011).

Finally, studies on CDS to hearing-impaired children with cochlear implants or hearing aids show that many features are more like speech directed at experience-matched, rather than age-matched, controls, showing caregiver accommodation to the needs of the child. For example, when addressing hearing-impaired children, caregivers exaggerate prosody more than when addressing age-matched controls (Bergeson et al. 2006). They may also adjust syntactic complexity in line with the child's linguistic experience and fine-tune it to the child's specific lexical knowledge (Odijk and Gillis 2021). So far, the evidence suggests that, although caregivers tend to accommodate children's hearing status, there is still considerable variation in the speech directed to hearing-impaired children (Arjmandi et al. 2022).

Overall, the study of CDS produced for children with clinical conditions including ASD, Down syndrome, visual impairment, hearing impairment, and more can yield new insights into how accommodation and responsiveness to children interact with specific beliefs and attitudes about children's capabilities. The main caveat associated with this research strand is that most studies have been conducted in Western industrialized countries. Because disability is a socially constructed category (Scheer and Groce 1988), the parental response to it must be understood in its cultural context, making it difficult to disentangle the effects of cultural expectations about these children from the effects of caregivers' capacity for accommodating them.

In sum, much more work is needed to understand how language transmission and language development proceed under such immense variability in conditions across and within societies, both in terms of how they are supported by CDS but also by the

9780192886491_Book.indb 92 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:28

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

wide variety of other linguistic and communicative behaviours children encounter and engage with in their social environments.

5.3 Studying language learning outcomes from CDS

Even though CDS only forms one part of the language surrounding children, it is canonically considered the part that is most important for language development, at least in some industrialized societies. Consequently, much research has focused on identifying which of its features can facilitate the aspects of language learning that are most strongly associated with social and economic success in such societies (e.g. boosting vocabulary size and literacy). Studies that examine the effect of CDS on language development can be divided into three major strands: (a) examining children's preference for, or orientation to, CDS compared to ADS; (b) testing the association between certain features of CDS and learning outcomes; and (c) investigating what features of CDS aid certain aspects of language learning over and above what is learnable from ADS. Studies of the first and second strand tend to measure outcomes directly in children. Studies of the third strand require experiments, often with adult participants, since a natural control of children learning language from ADS does not exist. All three strands also use computational models to estimate learnability.

5.3.1 Studying infant preference for CDS

The first study to experimentally investigate the preference of pre-linguistic infants for CDS (labelled infant-directed speech, 'IDS', in this literature) used the operant auditory preference procedure where infants are rewarded with a speech signal for turning their head to one or the other speaker mounted to each side of their seat (Fernald 1985). Four-month-old infants turned their heads longer in the direction of CDS showing that it captured their attention more than ADS. Subsequent studies tested infants from as young as 2 days (Cooper and Aslin 1990) and up until and even beyond 1 year of age, using a variety of procedures. While some studies presented naturalistic recordings of CDS and ADS, others controlled the distinguishing prosodic features, like elevated pitch and exaggerated pitch range. A meta-analysis of thirty-four studies (Dunst et al. 2012) confirmed a general preference for CDS with a medium-to-large effect size.

Infant studies are notorious for their small sample sizes which are liable to false positives. To rectify this, a cross-linguistic multilab consortium study aimed to replicate the CDS preference in 2,329 infants aged 3–15 months acquiring eleven different languages using the three most common procedures to assess infant attention: the single-screen central fixation procedure, the head-turn preference procedure as well

9780192886491_Book.indb 93 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:28

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

as eye-tracking (ManyBabies Consortium 2020). The findings confirmed a general infant preference for CDS, although the effect size was in the small-to-medium range, i.e. somewhat smaller than estimated by Dunst et al.'s (2012) meta-analytic estimate. It also demonstrated that the CDS preference increases with age, is strongest for a familiar language, and when tested with the head-turn preference procedure. These behavioural findings are complemented by infant neuroimaging studies, using event-related potentials (e.g. Zangl and Mills 2007) and functional near-infrared spectroscopy (e.g. Saito et al. 2007; Sulpizio et al. 2018) to uncover the patterns of brain activation that underpin the attentional processes responsible for the CDS preference.

The meta-analytic and multilab evidence shows that the attention-capturing effect of CDS prosody appears to be one of the most robust phenomena in child language development research. However, very few studies have explored whether other features beyond speech prosody, e.g. infant-directed action aka 'motionese' (Brand and Shallcross 2008), may indicate a wider preference for infant-directed communication. A related, albeit smaller research strand is investigating infants' ability to discriminate CDS features that signal different emotions or communicative intentions of caregivers (e.g. Mastropieri and Turkewitz 1999). Taken together, this research sheds light on the early processing capacities that seem to accord CDS a privileged status as a source of input for language development—children like to attend to it. Further studies into CDS preference in non-industrialized societies and societies with low prevalence of CDS are needed to test the universality of this phenomenon as a prerequisite for language transmission.

5.3.2 Linking features of CDS to learning outcomes

Many of the studies discussed in section 5.1 interpret the features that distinguish CDS from ADS as proxies for how easy it may be to learn certain aspects of language. Here we discuss studies that investigate the effects of CDS on child language learning directly. It is fair to say that this research strand comprises the largest proportion of published articles on CDS. It is correlational in nature, linking quantity and quality of CDS features to learning outcomes. Often, aggregate measures from CDS corpora obtained at one point in time are correlated with child learning outcomes at a later point in time (but see Slone et al. 2023 for recent work on how not only frequency but distribution of lexical items over time can affect learning).

Learning outcomes have been studied for a wide range of linguistic input domains, e.g. quantity of certain lexical items, constructions, morphosyntactic, pragmatic, and discourse features (e.g. Newport et al. 2020; Genovese et al. 2020; Davies et al. 2020), or distributional characteristics of CDS including extralinguistic features like caregiver gestures (e.g. O'Neill et al. 2005; Gogate et al. 2015), reciprocity, and turn-taking behaviour (e.g. Bedrosian et al. 1988; Kondaurova et al. 2020). The range of tested language learning outcome variables is equally broad and includes, but is not limited to, frequency and nature of infant vocalizations (e.g. Henning et al. 2005), infant word segmentation

9780192886491_Book.indb 94 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:28

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

ability (e.g. Thiessen et al. 1996), specific aspects of prosodic (e.g. McRoberts and Best 1997), phonological (e.g. Sim and Post 2022), morphosyntactic (e.g. Newport et al. 1977), or lexical development (e.g. Weizman and Snow 2001), development of broader discourse and turn-taking abilities (e.g. Bedrosian et al. 1988; Kondaurova et al. 2020) or composite measures of receptive and expressive language using various standardized tests (e.g. Wang et al. 2020). Of particular importance here are longitudinal studies that attempt to capture how the nature of associations between CDS features and learning outcomes changes with the child's age as both the aims of child-directed communication as well as children's cognitive capacities change (e.g. Rowe 2012). Summarizing the results of this expansive line of research in detail is beyond the scope of this chapter but a common conclusion is that CDS entails simplifications in many of these input domains that can facilitate language development, although not all simplifications are facilitators of language learning (Newport et al. 2020).

A smaller but equally important set of studies employs quantitative measures or computational modelling and machine learning to test specific theories about the nature of child language learning, using various learning algorithms that vary in the degree of implementation of bottom-up (unsupervised) vs. top-down (supervised) learning processes (e.g. Gaussian mixture models, neural networks, Bayesian models, decision trees). These approaches are invaluable for disentangling effects of different input features and their interactions on language learning. For example, several studies have evaluated the informativeness of various distributional patterns of word co-occurrences in CDS for learning of syntactic categories. Classification accuracy of three-word co-occurrence patterns, i.e. word frames like '*I see it*', '*I want it*', '*I got it*', instantiating non-adjacent dependencies ('AXB', where X is the target word to be classified as noun or verb) was compared with other patterns (e.g. 'XAB' or 'ABX') across several languages. The main results are that frequent *word* frames are informative for learning grammatical categories in some languages, like English, but *morpheme* frames (e.g. Russian: *ot slona* [from elephant.GEN], *k slonu* [towards elephant.DAT], *pod slonom* [underneath elephant.INSTR] seem to be more informative cross-linguistically, especially in morphologically rich languages like Russian or Turkish (e.g. Mintz 2003; Chemla et al. 2009; Moran et al. 2018). Other examples involve the use of distributional learning algorithms like word embeddings to demonstrate that causal meaning can be inferred more easily from CDS without recourse to syntax (You et al. 2021) or the use of Gaussian mixture-models and Expectation Maximization to show superior learnability of vowel categories from vowels that are prosodically exaggerated, as in CDS, compared to vowels that are prosodically unfocused, as in ADS (Adriaans and Swingley 2017). The latter study exemplifies the role of CDS in cross-domain facilitation of learning, e.g. when syntactic units are identified from the distribution of prosodic features, a phenomenon termed 'prosodic bootstrapping' (Morgan and Demuth 2014).

To understand how cultural differences or differences in SES affect language learning trajectories, correlational studies of this kind have directly examined the effect of variation in quantitative and qualitative features of CDS on language development. For example, Casillas et al. (2020) demonstrated that rural Tseltal Mayan children—who

9780192886491_Book.indb 95 /12_first_proofs/first_pr…/xml_for_typesetting 27-Mar-25 23:02:28

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

are typically raised in a non-child-centred manner—reach early linguistic milestones like babbling, first words, and first word combinations at roughly the same time as their middle-class Western counterparts. With respect to effects of CDS variation by SES, it has been shown that greater vocabulary diversity and utterance length in maternal CDS (and not SES *per se*) relates to earlier and faster vocabulary growth for US children and that adherence to language teaching practices can mediate the effect of maternal education on language development in Chinese children (Hoff 2003; Hoff and Tian 2005).

5.3.3 Comparing learning gains between CDS and ADS

Links between quantitative and qualitative features of CDS and language learning outcomes in children illuminate the mechanisms by which children process language input. They cannot, however, provide insight into whether CDS is specifically tailored to the task of language learning, as has frequently been claimed (e.g. Ferguson 1964; Saint-Georges et al. 2013; Golinkoff et al. 2015). For that, comparisons with appropriate ADS control conditions are needed to measure potential learning gains from CDS over and above those from ADS. ADS controls are infeasible to establish in naturalistic observations, as children cannot be randomly assigned to grow up under conditions where they are addressed in the same way as adults. This line of research is therefore dominated by experimental and modelling studies.

The main aim of the experimental studies is to examine whether certain features of CDS facilitate processing of some vital aspect of language learning. Using both adult and child participants, such experiments have tested questions like whether ostensive cues accompanying CDS result in more infant gaze following (Senju and Csibra 2008), whether presence of CDS-typical lexical modifications like reduplication (e.g. *choo-choo*, *wee-wee*) or diminutivization (e.g. *kitty*, *bunny*) result in better learning of inflectional morphology (e.g. Ševa et al. 2007) or of novel words (Ota and Skarabela 2016), or whether CDS prosody improves detection of syntactic constituents (Nelson et al. 1989), to name a few. Some of the experiments tested perception while others examined learning from limited input over the very short time frames that can be implemented in laboratory experiments. For the most part, such experiments have established processing and learning gains associated with the CDS feature in question thereby providing proof of concept for specific learning advantages from certain CDS features. However, they cannot establish the magnitude of gains that could emerge during prolonged learning in real-life adult–child interactions.

A more direct test of learning gains from CDS is possible using quantitative estimates and computational models to compare learning from large-scale naturalistic ADS and CDS corpora. These studies, which emerged within the last decade, are starting to paint a more complex and nuanced picture of what specific kinds of learning benefits CDS can have, over and above ADS, in each cultural context. For example, discriminability of phonemes in CDS is enhanced in some languages (e.g. in English CDS; Eaves et al. 2016) and reduced in others (e.g. in Japanese CDS; Martin et al. 2015). Crucially, the

9780192886491_Book.indb 96 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:28

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

computational approach makes it possible to examine the contribution of various types of information like acoustic, phonological, or phonotactic cues to different aspects of learning like word-form learning (Guevara-Rukoz et al. 2018) or word segmentation (Ludusan et al. 2021, 2022). It also makes it possible to investigate the suitability of different learning algorithms to capture this information (Ludusan et al. 2016, 2022). Computational studies permit not just quantification of the magnitude of the CDS benefit, but can also assess how much could be learned about that same aspect of language based on ADS. For example, Kodner (2019) showed considerable overlap between CDS and various genres of ADS in the prevalence of some structural properties of verbs (e.g. irregular verbs, direct-object/to-dative alternator verbs) among the most frequent verb types, suggesting that CDS and ADS may share many high type-frequency lexical items. As a result, a learning algorithm based on the Tolerance Principle (Yang 2016), implemented to simulate English past tense *-ed* learning, achieved similar outcomes with CDS and ADS. Similarly, Ludusan et al. (2016) showed that while prosodic boundaries were easier to detect from CDS input, they could also be learned from ADS input, albeit with slightly reduced outcomes, a finding that echoes results from laboratory experiments where learning of inflectional morphology was supported by diminutivization but not impossible without (Kempe and Brooks 2005). The latter point is important for gaining a better understanding of how language transmission can proceed despite relatively low prevalence of a distinct CDS register in some communities.

5.4 Explaining the drivers and functions of CDS

What drives the observed differences between CDS and ADS? Ferguson's (1964) original proposal, that CDS serves to aid language acquisition, has had considerable influence, so much so that in the field of language development research, a facilitative function of CDS is often taken for granted. Yet, surprisingly little work has addressed the proximate drivers of CDS. If CDS serves to aid language development, should we then consider it a form of teaching? If so, does this imply awareness of, and intent to teach the child, how to speak on the part of the caregiver? It has indeed been suggested that CDS may be consistent with pedagogical input sampling (Shafto and Goodman 2008) potentially rendering CDS a tool for teaching some aspects of language (Eaves et al. 2016). Are there mechanisms by which a teaching function may manifest without the intent to teach? Recent theorizing on the evolutionary origins of teaching behaviour in general may provide some answers: Brandl et al. (2023) have suggested that the human capacity for teaching is a cognitive gadget—a culturally transmitted behaviour that utilizes existing domain-general mechanisms in the service of an outcome that is adaptive under certain ecological, cultural, and social conditions. It seems reasonable

9780192886491_Book.indb 97 /12_first_proofs/first_pr[illegible]/xml_for_typesetting 27-Mar-25 23:02:28

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

to apply the same logic to CDS and to view it also as a culturally transmitted cognitive gadget.

As a cognitive gadget, CDS may capitalize on two domain-general mechanisms: affect expression and speech/communication accommodation. Specifically, the prosodic features of CDS may be expressions of the positive affect elicited by infants in their caregivers (e.g. Singh et al. 2002). Indeed, neuroimaging data demonstrate that, for mothers, exposure to their own infants is one of the most powerful positive mood inducers (Nitschke et al. 2004). In addition, caregivers may use the signalling function of exaggerated prosody to try to capture children's attention, which, in turn, may facilitate joint attention between child and caregiver, making each person's intended meanings more apparent to the other. When the attention-capturing prosody of CDS is attenuated, as it happens in depressed mothers, infant learning appears to proceed differently (Kaplan et al. 1999). Not only does CDS capture children's attention, but it can also regulate their affect and arousal, as evidenced by infants' ability to distinguish different intonation contours (e.g. Papousek et al. 1990). Furthermore, many features of CDS arise in response to children's vocal behaviour (Elmlinger et al. 2019) as rewards for the child but also as a form of speech accommodation on the part of the caregiver, like when raising one's pitch during interaction with an interlocutor who is emitting high-pitched vocalizations (Ko et al. 2016). Even though there is debate regarding whether alignment and speech accommodation in general are based on automatic processes like priming (Pickering and Garrod 2004) or involve effortful perspective taking (Epley et al. 2004), the cognitive gadget perspective on CDS would suggest that language evolution may not just be shaped by human cognitive capacity constraints (Kirby et al. 2008) but also by the effects of capabilities like affect expression and Theory of Mind on language transmission.

Conceptualizing CDS as a cognitive gadget that builds on the fundamental human capacities of affect expression and speech/communication accommodation does not preclude its use as an intentional teaching device under some circumstances. It is, however, unlikely that intentional teaching through CDS constitutes a human universal nor that language evolution is dependent on it. Instead, just like teaching of other cultural traits, intentional use of CDS to teach language may arise under conditions in which the lifetime and population productivity gain from being taught exceeds the productivity loss associated with teaching (Gurven et al. 2020). This may well be the case in many industrialized literacy-intensive societies where linguistic prowess contributes to individuals' lifetime productivity. Future research, including modelling work, will have to try to map out under what conditions cultural transmission processes may have combined, consolidated, and directed these domain-general capacities towards aiding the development of language and language-related skills (e.g. literacy).

A cultural evolution perspective on CDS as a culturally transmitted cognitive gadget can also help us gain a deeper understanding of how variation of CDS across cultures and social strata interacts with humans' highly robust ability to learn language: for millennia, many thousands of languages have been transmitted across generations in extraordinarily diverse environments. The evidence to date makes it clear that there

9780192886491_Book.indb 98 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:28

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

is not just one way in which language can be learned, nor one set of functions that it can serve; language allows humans to engage in many different types of encultured behaviours. Yet, in most industrialized and developing countries, research emphasis is often placed on vocabulary development as a prerequisite for literacy and STEM education. This emphasis elevates lexically diverse, standardized, child-centred speech to a gold-standard for language input and perpetuates societal expectations about the nature of 'optimal' CDS—expectations which are culturally, societally, and institutionally transmitted. A by-product of such standardization efforts is the marginalization of underpowered and underrepresented linguistic communities and their language transmission practices. Concomitantly, many efforts to boost children's early vocabulary development through interventions lack an awareness of, and engagement with, the broader societal pressures that strongly reinforce the stark inequalities in socioeconomic conditions they are attempting to overcome (see Rowe and Weisleder 2020 for a more in-depth review and Chater and Loewenstein 2023 for a broader perspective on the inadequacy of individual psychological solutions to societal problems).

5.5 Future directions

Over the past six decades, the emphasis of research has shifted from a characterization of CDS as the main source of child language input to a more nuanced understanding of CDS as just one of many different aspects of child-surrounding language and communication. Trying to capture how the richness of interactions with children enables development of language and other cognitive and social capacities is increasingly informed by broader theoretical perspectives put forth by bioecology (e.g. Bronfenbrenner 1979), non-linear dynamics system theory (Thelen and Smith 1994), evolutionary psychology (Csibra and Gergely 2009), and cultural evolution research (Creanza et al. 2017; Brandl et al. 2023). These perspectives acknowledge the multisystemic and multifunctional nature of communication that children are involved in and provide avenues for contextualizing the diversity of paths to language that children may take and, consequently, the different conditions under which language transmission proceeds. Specifically, we suggest that future research on child-directed and child-surrounding communication should consider the following research directions:

1. A first move will be to shift away from viewing child-directed language merely as a source of data from which children extract statistical and referential information, and to instead consider it part of a broader repertoire of culturally transmitted child-rearing behaviours and socialization activities. Doing so will provide deeper and more nuanced insights into how different proximate mechanisms (i.e. caregiver affective signalling, child affect regulation, speech/communication accommodation, or intentional teaching) interact to generate the language children are exposed to. Currently, our understanding of the different input types and different

9780192886491_Book.indb 99 /12_first_proofs/first_pr…/xml_for_typesetting 27-Mar-25 23:02:28

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

facets of learning is still limited. The cross-cultural work conducted to date is telling us that we are missing crucial information about different types of communication that includes the child, including talk directed to other (sometimes older) children and talk in which the child is a secondary addressee. Such a shift in emphasis can help to uncover the processes that mediate how ecology and culture shape the communication that children partake in. A cultural-evolution perspective may prove especially useful here: we can think of the different strategies employed in child-directed communication as culturally transmitted traits that encompass conventionalized templates for how to conduct interactions with children. Such templates can then be modified flexibly according to the needs of the interlocutors and the constraints of the situational context. This view shifts the explanatory burden of how child-directed communication supports language transmission away from constraints acting on the individual towards communicative strategies that are themselves transmitted as adaptive behaviours in various cultural and ecological contexts. This, in turn, will offer a broader perspective on the multiple constraints that may act on language evolution.

2. The nature of communication that is directed at the child, surrounds the child, or is relevant to the child more broadly needs to be viewed in the context of the richness of human activities and interactions in general. To date, a disproportionate emphasis on play with toys and picture-book reading in laboratory studies and on dyadic mother–child interactions may have provided a distorted picture that misses the diversity of communicative situations in which infants and young children acquire language. Moreover, comparison of child-directed to adult-directed communication is often restricted to a few standardized adult interactions (e.g. interviews between mothers and experimenters) which are of limited ecological validity and do not even come close to capturing the broad variety of adult communication (Weinstein and Baldwin 2022). In addition, the existing comparisons between CDS and ADS tend to focus on a few domains of language, mainly prosody and phonology, at the expense of others, including morphosyntactic, lexico-semantic, and pragmatic features. Considering a broader range of linguistic features, social situations, and interlocutors will provide a better understanding of what is truly unique about interactions with children and what is drawn from the broad and flexible repertoire of general human communication strategies. Of specific importance will be a move away from aggregate measures of CDS features towards a greater emphasis on how these features change dynamically across different social and physical activities (Casillas 2023).
3. A deeper appreciation of the multimodal nature of interactions that are relevant to the child will provide a better understanding of developmental cascades that support language development in the context of the development of other capacities. Expanding the focus on other modalities (e.g. gesture, gaze, touch) will shed light on how the different proximate functions of child-directed communication manifest themselves across different interaction channels. For example, in some cultures and situations, caregivers' affect signalling and regulation of the child's affective states

9780192886491_Book.indb 100 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:28

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

may be achieved through tactile rather than verbal interaction (e.g. Abu-Zhaya et al. 2017; de León 2021). The same may hold for how joint attention is initiated—there may be differences with respect to whether the primary channel is verbal, gestural, or tactile (Abels 2020). Moreover, depending on culture, intentional teaching efforts that involve scaffolding and modifying features of the environment may be directed towards skills other than language development (e.g. motor development, kinship knowledge; Karasik et al. 2015; Casillas 2023). Capturing the interplay and timing of interactions across modalities, domains, and skills will provide a fuller picture through which to appraise the role and relative importance of child-directed communication, and language transmission, in human development. This will entail supplementing data-gathering methodologies focusing on the auditory modality with those that include information on the role of visual, tactile, and kinesthetic interactions as components of the entire embodied relationship between caregivers and children (Shai and Belsky 2017).

4. The broadening of research perspectives advocated above necessitates a broadening of our methodological repertoire and integration of observational and experimental methods (Cristia et al. 2023). This will entail expanding and augmenting current analysis methods of long-form recordings to *in-vivo* approaches that capture contextual information (e.g. including logging of situational context alongside transcription of linguistic content) but also broadening elicitation methods used in the laboratory to include diverse situations and interlocutors. It will also require adapting experimental procedures in ways that ensure their relevance to specific cultural contexts. Crucially, to expand the cross-cultural evidence base we need to continue along the path of consortium-based developmental science such as that pioneered by the ManyBabies projects (e.g. ManyBabies Consortium 2020), but in a way that includes local researchers and underrepresented communities in the co-design of studies to adequately examine the wide variety of children's paths into language.

Broadening perspectives and research approaches to view child-directed communication as embedded in diverse social contexts has the potential to offer a deeper understanding not just of the many paths to language development but also of how cultural transmission of language, and ultimately, its evolution, are shaped by the range of conditions that characterize human social diversity.

References

Abels, M. (2020). 'Triadic interaction and gestural communication: Hierarchical and child-centered interactions of rural and urban Gujarati (Indian) caregivers and 9-month-old infants', *Developmental Psychology* 56(10): 1817.

Abu-Zhaya, R., Seidl, A., and Cristia, A. (2017). 'Multimodal infant-directed communication: How caregivers combine tactile and linguistic cues', *Journal of Child Language* 44(5): 1088–1116.

9780192886491_Book.indb 101 /12_first_proofs/first_pr/xml_for_typesetting 27-Mar-25 23:02:28

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Adi-Bensaid, L., Ben-David, A., and Tubul-Lavy, G. (2015). 'Content words in Hebrew child-directed speech', *Infant Behavior and Development* 40: 231–241.

Adriaans, F. and Swingley, D. (2017). 'Prosodic exaggeration within infant-directed speech: Consequences for vowel learnability', *The Journal of the Acoustical Society of America* 141(5): 3070–3078.

Akhtar, N. (2005). 'The robustness of learning through overhearing', *Developmental Science* 8(2): 199–209.

Arjmandi, M. K., Houston, D., and Dilley, L. C. (2022). 'Variability in quantity and quality of early linguistic experience in children with cochlear implants: Evidence from analysis of natural auditory environments', *Ear and Hearing* 43(2): 685–698.

Baran, J. A., Laufer, M. Z., and Daniloff, R. (1977). 'Phonological contrastivity in conversation: A comparative study of voice onset time', *Journal of Phonetics* 5(4): 339–350.

Bedrosian, J. L., Wanska, S. K., Sykes, K. M., et al. (1988). 'Conversational turn-taking violations in mother-child interaction', *Journal of Speech, Language, and Hearing Research* 31(1): 81–86.

Bentenuto, A., Perzolli, S., Bertamini, G., et al. (2021). 'Paternal speech directed to young children with Autism Spectrum Disorders and typical development', *Research in Developmental Disabilities* 112: 103886.

Bergelson, E., Amatuni, A., Dailey, S., et al. (2019). 'Day by day, hour by hour: Naturalistic language input to infants', *Developmental Science* 22(1): e12715.

Bergelson, E., Soderstrom, M., Schwarz, I. C., et al. (2023). 'Everyday language input and production in 1001 children from 6 continents', *PNAS* 120(52): e2300671120.

Bergeson, T. R., Miller, R. J., and McCune, K. (2006). 'Mothers' speech to hearing-impaired infants and children with cochlear implants', *Infancy* 10(3): 221–240.

Boersma, P. and Weenink, D. (2005). Praat: Doing phonetics by computer [computer program], http://www.praat.org/.

Brand, R. J. and Shallcross, W. L. (2008). 'Infants prefer motionese to adult-directed action', *Developmental Science* 11(6): 853–861.

Brandl, E., Mace, R., and Heyes, C. (2023). 'The cultural evolution of teaching', *Evolutionary Human Sciences* 5: e14.

Bronfenbrenner, U. (1979). *The Ecology of Human Development: Experiments by Nature and Design*. Cambridge, MA: Harvard University Press.

Bryant, G. A. and Barrett, H. C. (2007). 'Recognizing intentions in infant-directed speech: Evidence for universals', *Psychological Science* 18(8): 746–751.

Bunce, J., Soderstrom, M., Bergelson, E., et al. (2024). 'A cross-linguistic examination of young children's everyday language experiences', *Journal of Child Language* 1–29.

Burnham, D., Kitamura, C., and Vollmer-Conna, U. (2002). 'What's new, pussycat? On talking to babies and animals', *Science* 296(5572): 1435–1435.

Campbell, E. E. and Bergelson, E. (2022). 'Making sense of sensory language: Acquisition of sensory knowledge by individuals with congenital sensory impairments', *Neuropsychologia* 174: 108320.

Casey, K., Elliott, M., Mickiewicz, E., et al. (2022). 'Sticks, leaves, buckets, and bowls: Distributional patterns of children's at-home object handling in two subsistence societies', in J. Culbertson, A. Perfors, H. Rabagliati, and V. Ramenzoni (eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society*, 927–933.

Casillas, M. (2023). 'Learning language in vivo', *Child Development Perspectives* 17(1): 10–17.

Casillas, M., Brown, P., and Levinson, S. C. (2020). 'Early language experience in a Tseltal Mayan village', *Child Development* 91(5): 1819–1835.

9780192886491_Book.indb 102 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:28

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Casillas, M., Brown, P., and Levinson, S. C. (2021). 'Early language experience in a Papuan community', *Journal of Child Language* 48(4): 792–814.

Chater, N. and Loewenstein, G. (2023). 'The i-frame and the s-frame: How focusing on individual-level solutions has led behavioral public policy astray', *Behavioral and Brain Sciences* 46: e147.

Chemla, E., Mintz, T. H., Bernal, S., and Christophe, A. (2009). 'Categorizing words using 'frequent frames': What cross-linguistic analyses reveal about distributional acquisition strategies', *Developmental Science* 12(3): 396–406.

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.

Christiansen, M. H. and Chater, N. (2008). 'Language as shaped by the brain', *Behavioral and Brain Sciences* 31(5): 489–509.

Cooper, R. P. and Aslin, R. N. (1990). 'Preference for infant-directed speech in the first month after birth', *Child Development* 61(5): 1584–1595.

Cox, C., Bergmann, C., Fowler, E., et al. (2022). 'A systematic review and Bayesian meta-analysis of the acoustic features of infant-directed speech', *Nature Human Behaviour* 7(1): 114–133.

Creanza, N., Kolodny, O., and Feldman, M. W. (2017). 'Cultural evolutionary theory: How culture evolves and why it matters', *Proceedings of the National Academy of Sciences* 114(30): 7782–7789.

Cristia, A. (2022). 'A systematic review suggests marked differences in the prevalence of infant-directed vocalization across groups of populations', *Developmental Science* e13265.

Cristia, A. and Seidl, A. (2014). 'The hyperarticulation hypothesis of infant-directed speech', *Journal of Child Language* 41(4): 913–934.

Cristia, A., Dupoux, E., Gurven, M., and Stieglitz, J. (2019). 'Child-directed speech is infrequent in a forager-farmer population: A time allocation study', *Child Development* 90(3): 759–773.

Cristia, A., Foushee, R., Aravena-Bravo, P., et al. (2023). 'Combining observational and experimental approaches to the development of language and communication in rural samples: Opportunities and challenges', *Journal of Child Language* 50(3): 495–517.

Csibra, G. and Gergely, G. (2009). 'Natural pedagogy', *Trends in Cognitive Sciences* 13(4): 148–153.

Curcio, F. and Paccia, J. (1987). 'Conversations with autistic children: Contingent relationships between features of adult input and children's response adequacy', *Journal of Autism and Developmental Disorders* 17(1): 81–93.

Cychosz, M., Villanueva, A., and Weisleder, A. (2021). 'Efficient estimation of children's language exposure in two bilingual communities', *Journal of Speech, Language and Hearing Research* 64(10): 3843–3866.

Dailey, S. and Bergelson, E. (2022). 'Language input to infants of different socioeconomic statuses: A quantitative meta-analysis', *Developmental Science* 25(3): e13192.

Davies, C., Lingwood, J., and Arunachalam, S. (2020). 'Adjective forms and functions in British English child-directed speech', *Journal of Child Language* 47(1): 159–185.

de Falco, S., Venuti, P., Esposito, G., and Bornstein, M. H. (2011). 'Maternal and paternal pragmatic speech directed to young children with Down syndrome and typical development', *Infant Behavior and Development* 34(1): 161–169.

de León, L. (2021). 'The soothing nursing niche: Affective touch, talk, and pragmatic responses to Mayan infants' crying', *Journal of Pragmatics* 185: 146–160.

Dunst, C., Gorman, E., and Hamby, D. (2012). 'Preference for infant-directed speech in preverbal young children', *Center for Early Literacy Learning* 5(1): 1–13.

9780192886491_Book.indb 103 /12_first_proofs/first_pr[illegible]/xml_for_typesetting 27-Mar-25 23:02:28

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Eaves Jr, B. S., Feldman, N. H., Griffiths, T. L., and Shafto, P. (2016). 'Infant-directed speech is consistent with teaching', *Psychological Review* 123(6): 758.

Elmlinger, S. L., Schwade, J. A., and Goldstein, M. H. (2019). 'The ecology of prelinguistic vocal learning: Parents simplify the structure of their speech in response to babbling', *Journal of Child Language* 46(5): 998–1011.

Englund, K. T. (2005). 'Voice onset time in infant directed speech over the first six months', *First Language* 25(2): 219–234.

Epley, N., Keysar, B., Van Boven, L., and Gilovich, T. (2004). 'Perspective taking as egocentric anchoring and adjustment', *Journal of Personality and Social Psychology* 87(3): 327.

Ferguson, C. (1977). 'Baby talk as a simplified register', in C. Snow and C. Ferguson (eds.), *Talking to Children*, 219–236. Cambridge: Cambridge University Press.

Ferguson, C. A. (1964). 'Baby talk in six languages', *American Anthropologist* 66(6): 103–114.

Fernald, A. (1985). 'Four-month-old infants prefer to listen to motherese', *Infant Behavior and Development* 8(2): 181–195.

Fernald, A. (1992). 'Human maternal vocalizations to infants as biologically relevant signals: An evolutionary perspective', in J. H. Barkow, L. Cosmides, and J. Tooby (eds.), *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, 391–449. New York: Oxford University Press.

Fernald, A., Taeschner, T., Dunn, J., et al. (1989). 'A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants', *Journal of Child Language* 16(3): 477–501.

Fidler, D. J. (2003). 'Parental vocalizations and perceived immaturity in Down syndrome', *American Journal on Mental Retardation* 108(6): 425–434.

Foushee, R., Horton, G., and Srinivasan, M. (2023). 'Capturing the complete language environment: What could children learn by 'listening in'?', *PsyArxiv* [preprint], https://doi.org/10.31234/osf.io/qu3ay.

Genovese, G., Spinelli, M., Lauro, L. J. R., et al. (2020). 'Infant-directed speech as a simplified but not simple register: A longitudinal study of lexical and syntactic features', *Journal of Child Language* 47(1): 22–44.

Gogate, L., Maganti, M., and Bahrick, L. E. (2015). 'Cross-cultural evidence for multimodal motherese: Asian Indian mothers' adaptive use of synchronous words and gestures', *Journal of Experimental Child Psychology* 129: 110–126.

Golinkoff, R. M., Can, D. D., Soderstrom, M., and Hirsh-Pasek, K. (2015). '(Baby) talk to me: The social context of infant-directed speech and its effects on early language acquisition', *Current Directions in Psychological Science* 24(5): 339–344.

Guevara-Rukoz, A., Cristia, A., Ludusan, B., et al. (2018). 'Are words easier to learn from infant- than adult-directed speech? A quantitative corpus-based investigation', *Cognitive Science* 42(5): 1586–1617.

Gurven, M. D., Davison, R. J., and Kraft, T. S. (2020). 'The optimal timing of teaching and learning across the life course', *Philosophical Transactions of the Royal Society B* 375(1803): 20190500.

Hart, B. and Risley, T. R. (1992). 'American parenting of language-learning children: Persisting differences in family-child interactions observed in natural home environments', *Developmental Psychology* 28(6): 1096.

Hart, B. and Risley, T. R. (2003). 'The early catastrophe', *Education Review* 17(1).

Hayes, D. P. and Ahrens, M. G. (1988). 'Vocabulary simplification for children: A special case of "motherese"?', *Journal of Child Language* 15(2): 395–410.

9780192886491_Book.indb 104 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:29

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Heath, S. B. (1983). *Ways with Words*. Cambridge: Cambridge University Press.
Henning, A., Striano, T., and Lieven, E. V. (2005). 'Maternal speech to infants at 1 and 3 months of age', *Infant Behavior and Development* 28(4): 519–536.
Henrich, J., Heine, S., and Norenzayan, A. (2010). 'The weirdest people in the world?', *Behavioral and Brain Sciences* 33: 61–83.
Hills, T. (2013). 'The company that words keep: Comparing the statistical structure of child- versus adult-directed language', *Journal of Child Language* 40(3): 586–604.
Hoff, E. (2003). 'The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech', *Child Development* 74(5): 1368–1378.
Hoff, E. and Tian, C. (2005). 'Socioeconomic status and cultural influences on language', *Journal of Communication Disorders* 38(4): 271–278.
Hoff-Ginsberg, E. (1991). 'Mother-child conversation in different social classes and communicative settings', *Child Development* 62(4): 782–796.
Holzrichter, A. S. and Meier, R. P. (2000). 'Child-directed signing in American sign language', in Ch. Chamberlain, J. P. Morford, and R. I. Mayberry (eds.), *Language Acquisition by Eye*, 25–40. Mahwah, NJ: Lawrence Erlbaum Associates.
Jones, S. (2007). ' "Opposites" in discourse: A comparison of antonym use across four domains'. *Journal of Pragmatics* 39(6): 1105–1119.
Kaplan, P. S., Bachorowski, J. A., and Zarlengo-Strouse, P. (1999). 'Child-directed speech produced by mothers with symptoms of depression fails to promote associative learning in 4-month-old infants', *Child Development* 70(3): 560–570.
Karasik, L. B., Tamis-LeMonda, C. S., Adolph, K. E., and Bornstein, M. H. (2015). 'Places and postures: A cross-cultural comparison of sitting in 5-month-olds', *Journal of Cross-Cultural Psychology* 46(8): 1023–1038.
Kempe, V. and Brooks, P. J. (2005). 'The role of diminutives in the acquisition of Russian gender: Can elements of child-directed speech aid in learning morphology?', *Language Learning* 55(S1): 139–176.
Kempe, V., Brooks, P. J., and Pirott, L. (2001). 'How can child-directed speech facilitate the acquisition of morphology', in M. Almgren, A. Barreña, M.-J. Ezeizabarrena, I. Idiazabal, and B. MacWhinney (eds.), *Research on Child Language Acquisition: Proceedings of the 8th Conference of the International Association for the Study of Child Language*, 1237–1247. Somerville, MA: Cascadilla Press.
Kemper, S. and Harden, T. (1999). 'Experimentally disentangling what's beneficial about elderspeak from what's not', *Psychology and Aging* 14(4): 656.
Kirby, S., Cornish, H., and Smith, K. (2008). 'Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language', *Proceedings of the National Academy of Sciences* 105(31): 10681–10686.
Ko, E. S., Seidl, A., Cristia, A., et al. (2016). 'Entrainment of prosody in the interaction of mothers with their young children', *Journal of Child Language* 43(2): 284–309.
Kodner, J. (2019). 'Estimating child linguistic experience from historical corpora', *Glossa: A Journal of General Linguistics* 4(1): 1–14.
Kondaurova, M. V., Smith, N. A., Zheng, Q., et al. (2020). 'Vocal turn-taking between mothers and their children with cochlear implants', *Ear and Hearing* 41(2): 362.
Konner, M. (1977). 'Infancy among the Kalahari desert San', in P. H. Leiderman (ed.), *Culture and Infancy: Variations in the Human Experience*, 287–328. San Diego, CA: Academic Press.

9780192886491_Book.indb 105 /12_first_proofs/first_pr[illegible]/xml_for_typesetting 27-Mar-25 23:02:29

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Konstantareas, M. M., Zajdeman, H., Homatidis, S., and McCabe, A. (1988). 'Maternal speech to verbal and higher functioning versus nonverbal and lower functioning autistic children', *Journal of Autism and Developmental Disorders* 18(4): 647–656.

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., et al. (1997). 'Cross-language analysis of phonetic units in language addressed to infants', *Science* 277(5326): 684–686.

Lancy, D. F. (2015). *The Anthropology of Childhood: Cherubs, Chattel, Changelings* (2nd ed.). Cambridge: Cambridge University Press.

Lavechin, M., Bousbibm, R., Bredin, H., et al. (2020). 'An open-source voice type classifier for child-centered daylong recordings', *Proceedings of Interspeech* 2020: 3072–3076.

Lester, N. A., Moran, S., Küntay, A. C., et al. (2022). 'Detecting structured repetition in child-surrounding speech: Evidence from maximally diverse languages', *Cognition* 221: 104986.

Lindblom, B. (1990). 'Explaining phonetic variation: A sketch of the H&H theory', in W. J. Hardcastle and A. Marchal (eds.), *Speech Production and Speech Modeling*, 403–39. Amsterdam: Kluwer Academic.

Liu, H., MacWhinney, B., Fromm, D., and Lanzi, A. (2023). 'Automation of language sample analysis', *Journal of Speech, Language, and Hearing Research* 66(7): 1–13.

Liu, H. M., Tsao, F. M., and Kuhl, P. K. (2009). 'Age-related changes in acoustic modifications of Mandarin maternal speech to preverbal infants and five-year-old children: A longitudinal study', *Journal of Child Language* 36(4): 909–922.

Ludusan, B., Cristia, A., Martin, A., et al. (2016). 'Learnability of prosodic boundaries: Is infant-directed speech easier?', *The Journal of the Acoustical Society of America* 140(2): 1239–1250.

Ludusan, B., Mazuka, R., and Dupoux, E. (2021). 'Does infant-directed speech help phonetic learning? A machine learning investigation', *Cognitive Science* 45(5): e12946.

Ludusan, B., Cristia, A., Mazuka, R., and Dupoux, E. (2022). 'How much does prosody help word segmentation? A simulation study on infant-directed speech', *Cognition* 219: 104961.

MacWhinney, B. and Snow, C. (1985). 'The child language data exchange system', *Journal of Child Language* 12(2): 271–295.

ManyBabies Consortium (2020). 'Quantifying sources of variability in infancy research using the infant-directed-speech preference', *Advances in Methods and Practices in Psychological Science* 3(1): 24–52.

Martin, A., Schatz, T., Versteegh, M., et al. (2015). 'Mothers speak less clearly to infants than to adults: A comprehensive test of the hyperarticulation hypothesis', *Psychological Science* 26(3): 341–347.

Masek, L. R., Ramirez, A. G., McMillan, B. T., et al. (2021). 'Beyond counting words: A paradigm shift for the study of language acquisition', *Child Development Perspectives* 15(4): 274–280.

Mastropieri, D. and Turkewitz, G. (1999). 'Prenatal experience and neonatal responsiveness to vocal expressions of emotion', *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology* 35(3): 204–214.

McMurray, B., Kovack-Lesh, K. A., Goodwin, D., and McEchron, W. (2013). 'Infant directed speech and the development of speech perception: Enhancing development or an unintended consequence?', *Cognition* 129(2): 362–378.

McRoberts, G. W., and Best, C. T. (1997). 'Accommodation in mean fo during mother–infant and father–infant vocal interactions: A longitudinal case study'. *Journal of Child Language* 24(3): 719–736.

Mintz, T. H. (2003). 'Frequent frames as a cue for grammatical categories in child directed speech', *Cognition* 90(1): 91–117.

9780192886491_Book.indb 106 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:29

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Montag, J. L. (2020). 'New insights from daylong audio transcripts of children's language environments', in S. Denson, M. Mack, Y. Xu, and B. C. Armstrong (eds.), *Proceedings of the 42nd Annual Virtual Meeting of the Cognitive Science Society*, 3005–3011.

Moran, S., Blasi, D. E., Schikowski, R., et al. (2018). 'A universal cue for grammatical categories in the input to children: Frequent frames', *Cognition* 175: 131–140.

Morgan, J. L. and Demuth, K. (eds.). (2014). *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition*. New York: Psychology Press.

Nelson, D. G. K., Hirsh-Pasek, K., Jusczyk, P. W., and Cassidy, K. W. (1989). 'How the prosodic cues in motherese might assist language learning', *Journal of Child Language* 16(1): 55–68.

Newport, E. L., Gleitman, H., and Gleitman, L. R. (1977). 'Mother, I'd rather do it myself: Some effects and non-effects of maternal speech style', in C. E. Snow and C. Ferguson (eds.), *Talking to Children*, 109–149. Cambridge: Cambridge University Press.

Newport, E. L., Gleitman, H., and Gleitman, L. R. (2020). 'Mother, I'd rather do it myself', in L. Gleitman and J. Lidz (eds.) *Sentence First, Arguments Afterward: Essays in Language and Learning*, 141. Oxford: Oxford University Press.

Nitschke, J. B., Nelson, E. E., Rusch, B. D., et al. (2004). 'Orbitofrontal cortex tracks positive mood in mothers viewing pictures of their newborn infants', *Neuroimage* 21(2): 583–592.

O'Neill, M., Bard, K. A., Linnell, M., and Fluck, M. (2005). 'Maternal gestures with 20-month-old infants in two contexts', *Developmental Science* 8(4): 352–359.

Ochs, E. (1982). 'Talking to children in Western Samoa', *Language in Society* 11: 77–104.

Odijk, L. and Gillis, S. (2021). 'Tailoring the input to children's needs: The use of fine lexical tuning in speech directed to normally hearing children and children with cochlear implants', *Frontiers in Psychology* 12: 2349.

Onnis, L., Esposito, G., Venuti, P., and Edelman, S. (2021). 'Parental speech to typical and atypical populations: A study on linguistic partial repetition', *Language Sciences* 83: 101311.

Oswalt, R. L. (1976). 'Baby talk and the genesis of some basic Pomo words', *International Journal of American Linguistics* 42(1): 1–13.

Ota, M. and Skarabela, B. (2016). 'Reduplicated words are easier to learn', *Language Learning and Development* 12(4): 380–397.

Papousek, M., Bornstein, M. H., Nuzzo, C., et al. (1990). 'Infant responses to prototypical melodic contours in parental speech', *Infant Behavior and Development* 13: 539–545.

Perez-Pereira, M. and Resches, M. (2008). 'Pragmatic analysis of conversations between blind children and their mothers and the issue of maternal directivity', *Infancia y Aprendizaje* 31(4): 425–447.

Piazza, G., Martin, C. D., and Kalashnikova, M. (2022). 'The acoustic features and didactic function of foreigner-directed speech: A scoping review', *Journal of Speech, Language, and Hearing Research* 65(8): 2896–2918.

Pickering, M. J. and Garrod, S. (2004). 'Toward a mechanistic psychology of dialogue', *Behavioral and Brain Sciences* 27(2): 169–190.

Pisani, S., Gautheron, L., and Cristia, A. (2021). 'Long-form recordings: From A to Z', DOI: 10.5281/zenodo.6685828

Pye, C. (1986). 'Quiché Mayan speech to children' *Journal of Child Language* 13(1): 85–100.

Räsänen, O., Seshadri, S., Lavechin, M., et al. (2021). 'ALICE: An open-source tool for *automatic* measurement of phoneme, syllable, and word counts from child-centered daylong recordings', *Behavior Research Methods* 53: 818–835.

9780192886491_Book.indb 107 /12_first_proofs/first_pr/xml_for_typesetting 27-Mar-25 23:02:29

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Rogoff, B., Mistry, J., Göncü, A., et al. (1993). 'Guided participation in cultural activity by toddlers and caregivers', *Monographs of the Society for Research in Child Development* 58(8): i–179.

Rowe, M. L. (2012). 'A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development', *Child Development* 83(5): 1762–1774.

Rowe, M. L. and Weisleder, A. (2020). 'Language development in context', *Annual Review of Developmental Psychology* 2: 201–223.

Saint-Georges, C., Chetouani, M., Cassel, R., et al. (2013). 'Motherese in interaction: At the cross-road of emotion and cognition? (A systematic review)', *PloS one* 8(10): e78103.

Saito, Y., Aoyama, S., Kondo, T., et al. (2007). 'Frontal cerebral blood flow change associated with infant-directed speech', *Archives of Disease in Childhood-Fetal and Neonatal Edition* 92(2): F113–F116.

Saxton, M. (2008). 'What's in a name? Coming to terms with the child's linguistic environment', *Journal of Child Language* 35(3): 677–686.

Scaff, C., Casillas, M., Stieglitz, J., and Cristia, A. (2022). 'Characterization of children's verbal input in a forager-farmer population using long-form audio recordings and diverse input definitions', *PsyArxiv* [preprint], https://doi.org/10.31234/osf.io/mt6nz.

Scheer, J. and Groce, N. (1988). 'Impairment as a human constant: Cross-cultural and historical perspectives on variation', *Journal of Social Issues* 44(1): 23–37.

Schieffelin, B. B. (1985). 'The acquisition of Kaluli', in D. I. Slobin (ed.), *The Crosslinguistic Study of Language Acquisition. The Data* (Vol. 1), 525–593. Hillsdale, NJ: Erlbaum.

Senju, A. and Csibra, G. (2008). 'Gaze following in human infants depends on communicative signals', *Current Biology* 18(9): 668–671.

Seva, N., Kempe, V., Brooks, P. J., et al. (2007). 'Crosslinguistic evidence for the diminutive advantage: Gender agreement in Russian and Serbian children', *Journal of Child Language* 34(1): 111–131.

Shafto, P. and Goodman, N. (2008). 'Teaching games: Statistical sampling assumptions for learning in pedagogical situations', in *Proceedings of the 30th annual conference of the Cognitive Science Society*, 1632–1637. Austin, TX: Cognitive Science Society.

Shai, D. and Belsky, J. (2017). 'Parental embodied mentalizing: How the nonverbal dance between parents and infants predicts children's socio-emotional functioning', *Attachment & Human Development* 19(2): 191–219.

Shneidman, L. A. and Goldin-Meadow, S. (2012). 'Language input and acquisition in a Mayan village: How important is directed speech?', *Developmental Science* 15(5): 659–673.

Sim, J. H. and Post, B. (2022). 'Variation in quality of maternal input and development of coda stops in English-speaking children in Singapore'. *Journal of Child Language* 49(6): 1147–1172.

Singh, L., Morgan, J. L., and Best, C. T. (2002). 'Infants' listening preferences: Baby talk or happy talk?', *Infancy* 3(3): 365–394.

Singh, L., Cristia, A., Karasik, L. B., et al. (2023). 'Diversity and representation in infant research: Barriers and bridges toward a globalized science of infant development', *Infancy* 28: 708–737.

Slone, L. K., Abney, D. H., Smith, L. B., and Yu, C. (2023). 'The temporal structure of parent talk to toddlers about objects', *Cognition* 230: 105266.

Sperry, D. E., Sperry, L. L., and Miller, P. J. (2019). 'Reexamining the verbal environments of children from different socioeconomic backgrounds', *Child Development* 90(4): 1303–1318.

9780192886491_Book.indb 108 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:29

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Sulpizio, S., Doi, H., Bornstein, M. H., et al. (2018). 'fNIRS reveals enhanced brain activation to female (versus male) infant directed speech (relative to adult directed speech) in Young Human Infants', *Infant Behavior and Development* 52: 89–96.

Sundberg, U. and Lacerda, F. (1999). 'Voice onset time in speech to infants and adults', *Phonetica* 56(3–4): 186–199.

Tal, S., Grossman, E., Rohde, H., and Arnon, I. (2023). 'Speakers use more redundant references with language learners: Evidence for communicatively-efficient referential choice', *Journal of Memory and Language* 128: 104378.

Tamis-LeMonda, C. S., Custode, S., Kuchirko, Y., et al. (2019). 'Routine language: Speech directed to infants during home activities', *Child Development* 90(6): 2135–2152.

Thelen, E. and Smith, L. B. (1994). *A Dynamic Systems Approach to the Development of Cognition and Action*. Cambridge, MA: MIT Press.

Thiessen, E. D., Hill, E. A., and Saffran, J. R. (2005). 'Infant-directed speech facilitates word segmentation', *Infancy* 7(1): 53–71.

Tubul-Lavy, G., Jokel, A., Leon-Attia, O., and Gabis, L. V. (2020). 'Content words in child-directed speech of mothers toward children with autism spectrum disorder', *American Journal of Speech-Language Pathology* 29(3): 1434–1447.

Vogt, P., Mastin, J. D., and Schots, D. M. A. (2015). 'Communicative intentions of child-directed speech in three different learning environments: Observations from the Netherlands, and rural and urban Mozambique', *First Language* 35(4–5): 341–358.

Walton, K. M. and Ingersoll, B. R. (2015). 'The influence of maternal language responsiveness on the expressive speech production of children with autism spectrum disorders: A micro-analysis of mother–child play interactions', *Autism* 19(4): 421–432.

Wang, Y., Williams, R., Dilley, L., and Houston, D. M. (2020). 'A meta-analysis of the predictability of LENA™ automated measures for child language development', *Developmental Review* 57: 100921.

Weinstein, N. and Baldwin, D. (2022). 'Reification of infant-directed speech? Exploring assumptions shaping infant-directed speech research', *Culture & Psychology* 1354067X221147683.

Weisleder, A. and Fernald, A. (2013). 'Talking to children matters: Early language experience strengthens processing and builds vocabulary', *Psychological Science* 24(11): 2143–2152.

Weizman, Z. O. and Snow, C. E. (2001). 'Lexical output as related to children's vocabulary acquisition: Effects of sophisticated exposure and support for meaning', *Developmental Psychology* 37(2): 265.

Wong, P. and Ng, K. W. S. (2018). 'Testing the hyperarticulation and prosodic hypotheses of child-directed speech: Insights from the perceptual and acoustic characteristics of child-directed Cantonese tones', *Journal of Speech, Language, and Hearing Research* 61(8): 1907–1925.

Yang, C. (2016). *The Price of Linguistic Productivity*. Cambridge, MA: MIT Press.

You, G., Bickel, B., Daum, M. M., and Stoll, S. (2021). 'Child-directed speech is optimized for syntax-free semantic inference', *Scientific Reports* 11(1): 16527.

Zangl, R. and Mills, D. L. (2007). 'Increased brain activity to infant-directed speech in 6-and 13-month-old infants', *Infancy* 11(1): 31–62.

Zebrowski, P. M., Weiss, A. L., Savelkoul, E. M., and Hammer, C. S. (1996). 'The effect of maternal rate reduction on the stuttering, speech rates and linguistic productions of children who stutter: Evidence from individual dyads', *Clinical Linguistics & Phonetics* 10(3): 189–206.

9780192886491_Book.indb 109 /12_first_proofs/first_pr…/xml_for_typesetting 27-Mar-25 23:02:29

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

9780192886491_Book.indb 110 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:29

CHAPTER 6

CORPUS-BASED APPROACHES TO EVOLUTIONARY DYNAMICS IN LANGUAGE

STEFAN HARTMANN AND MICHAEL PLEYER

6.1 Introduction

Corpus linguistics, i.e. the systematic quantitative and/or qualitative analysis of authentic language data, has become one of the most important approaches in contemporary linguistics (see e.g. Lüdeling and Kytö 2008; Paquot and Gries 2020). The claim that corpus data can prove insightful for studying the evolution of language might, at first glance, seem surprising. A corpus is a structured collection of authentic language data that usually aims at being representative for a specific language variety or a linguistic domain (see e.g. Kübler and Zinsmeister 2015: 2). As such, a crucial prerequisite for a corpus is that there are extant data for the language in question. Language evolution research, by contrast, is usually concerned with unravelling the origins of language, i.e. with the question of how the human capacity for language emerged and how fully-fledged human language may have developed from a hypothesized protolanguage stage that fills the gap between the non-linguistic communication systems of non-human animals and linguistic communication. But especially for the latter process, studying present-day language can arguably prove insightful. According to Bybee (2010: 202), 'there is every reason to suppose that the very first grammatical constructions emerged in the same way as those observed in more recent history' (see also Heine and Kuteva 2007; Arbib 2012; Hartmann and Pleyer 2021).

As such, historical language change has come to be considered an important source of evidence for language evolution research (see Pleyer and Hartmann 2024: ch. 6). While it is a matter of debate to what extent the very origins of linguistic communication can be explained by drawing on mechanisms observed in historical language change (see e.g. Mendívil-Giró 2019 for discussion), there is a growing consensus that

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

the study of cultural evolution is key to understanding how linguistic structure comes about. This becomes abundantly clear when we look at the topics represented at the pertinent conferences in the field: drawing on a quantitative analysis of almost 350 abstracts from 2006 to 2016, Bergmann and Dale (2016) show that topics related to cultural evolution and iterated learning (see Tamariz and Papa, Chapter 1, this volume) are central to the International Conference on the Evolution of Language (Evolang; for a follow-up analysis with similar results, see Wacewicz et al. 2022). In a similar vein, Hartmann et al. (2024) show that cultural evolution topics have been central to the 2022 Joint Conference on Language Evolution (JCoLE). While experimental methods such as artificial language learning (see Tal et al., Chapter 2, this volume) and communication games (Müller and Raviv, Chapter 3, this volume) are well suited for studying the cultural evolution of language, corpus studies complement such approaches in an important way as they offer insights based on authentic, naturalistic data.

In this chapter, we will first discuss terminological issues, focusing on the question of what belongs to the field of corpus linguistics and, relatedly, what a linguistic corpus is (section 6.2). Section 6.3 reviews a number of studies that exemplify the potential of corpus studies for gaining insights about the evolution of language. Section 6.4 then presents a more detailed case study of studies that have used historical corpus data to disentangle the role of stochastic drift and selection in language evolution. Section 6.5 concludes the chapter.

6.2 What is corpus linguistics and how does it work?

First and foremost, corpus linguistics is a specific research method that is used to address linguistic research questions in a systematic, empirically valid way. There is some debate as to whether corpus linguistics is also a paradigm or a discipline (see e.g. Taylor 2008), but in the present chapter we focus on corpus linguistics as a method in language evolution research.

Corpus linguistics draws on structured collections of authentic linguistic data, so-called corpora (see e.g. McEnery and Wilson 2001; Lemnitzer and Zinsmeister 2015). We deliberately choose a broad definition of *corpus* focusing on three features: (i) corpora are *structured* collections of data, i.e. they are compiled with regard to specific linguistic criteria, which means that e.g. the internet as a whole would not qualify as a corpus (Kilgarriff 2007), while corpora compiled from the web such as the 'Corpora from the Web' (Schäfer and Bildhauer 2012), WaCkY (Baroni et al. 2009), or the TenTen family of corpora (Jakubíček et al. 2013) would. The linguistic criteria that guide different data collection regimes differ depending on the purpose of the corpus and the specific research question that the corpus is supposed to address. A so-called reference corpus aims at being representative for the language or language variety it documents. For instance,

9780192886491_Book.indb 112 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:29

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

the widely used British National Corpus (BNC; BNC Consortium 2007) aims at being representative of spoken and written British English in the late twentieth century. Given that languages are not homogeneous entities but rather characterized by a broad array of varieties (e.g. dialects, regiolects, idiolects, but also registers, genres, and styles), representativeness is, of course, a problematic notion as it is impossible to come up with a complete list of 'categories' along which different varieties of a language differ, and that, therefore, would have to be taken into account when compiling a corpus that is fully representative of a given language (Hunston 2008: 161). As an approximation, corpora are usually *balanced* along a variety of pertinent criteria, e.g. text types such as newspaper or fiction in the case of written corpora, or age and gender in the case of spoken and signed corpora (see e.g. McEnery and Wilson 2001: 10–13). (ii) Corpora are collections of *authentic* linguistic data, as opposed to, say, introspectively invented example sentences. That being said, authenticity is of course a matter of degree. In most cases, linguistic corpora consist of data that were not created for the purpose of linguistic analysis—for instance, newspapers, novels, movie subtitles, or, in the case of multimodal corpora, TV shows. But especially spoken language corpora that document everyday conversations often consist of data that were explicitly collected for the purpose of linguistic analysis, with the informed consent of the interlocutors. This can, of course, lead to an observer effect (i.e. participants behave differently when they know that they are observed[1]). However, a number of strategies have been developed to minimize this problem (see e.g. Gut 2020). One field in which such corpora are indispensable is language acquisition research, as naturalistic child language data is much harder to obtain from extant sources than data documenting adult speech or writing (see e.g. Behrens 2008). Similarly, research on signed languages usually has to rely on elicited data as collected in sign language corpora (see e.g. Hodge and Goico 2022, who also discuss the concepts of *natural(istic)* vs. *elicited* data). (iii) Corpora are collections of *linguistic* data. That is, they document language use. The currently most widely used corpora do so in the written modality, but an increasing number of resources also document spoken and signed language. But some corpora do not only focus on strictly linguistic aspects in the narrow sense—a number of multimodal corpora put paraverbal and/or non-verbal information centre stage. These include prosodic and intonational aspects as well as co-speech gesture. Especially the latter aspect is also highly relevant from the perspective of language evolution research as the study of co-speech gesture can contribute to evaluating hypotheses regarding the question of whether language emerged in the vocal, or in the gestural, modality, or if it was multimodal from the outset (see e.g. Armstrong and Wilcox 2007; Perlman 2017; among many others).

The definition just offered covers a wide array of linguistic corpora. Some definitions that can be found in the literature are more narrow. For instance, some authors add as another criterion that a corpus has to be machine-readable (e.g. McEnery and Wilson

[1] Labov (1972: 209) calls this the Observer's Paradox: 'the aim of linguistic research in the community must be to find out how people talk when they are not being systematically observed; yet we can only obtain these data by systematic observation'.

9780192886491_Book.indb 113 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:29

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

2001: 1) or electronically available (Kübler and Zinsmeister 2015: 4). While this is a typical feature of a linguistic corpus in the digital age, we do not see it as a necessary criterion. In principle, a collection of, say, physical handwritten letters could serve as a corpus as well, and while there are many research-practical and -ethical reasons for making research data available in a digitized format and ideally transcribing them (ranging from easier usability to replicability and re usability), in principle, the analyst could also work with the original documents and obtain the same results they would arrive at when working with a digitized version. Thus, in our view, nothing speaks against treating physical documents as linguistic corpora, at least in theory. Another criterion proposed in some definitions is that they have to be large (Stefanowitsch 2020: 23). This criterion is related to the idea of representativeness: if we want to get a good impression of how a specific word or construction is used in a given language, our corpus must be large enough to cover different possible usage variants. As such, the required size of a corpus also depends on the research question: if we investigate low-frequency phenomena, we obviously need very large corpora to find enough attestations. For investigating high-frequency phenomena, and/or for purely qualitative analyses, smaller corpora may be sufficient. As such, size is definitely a criterion that is crucial in corpus design decisions, but a linguistic corpus does not necessarily have to be large (which is a very relative concept anyway).

In general, it is always important to reflect the representativeness and appropriateness of a corpus selected for a particular research question. For example, Pechenik et al. (2015) show that there are strong limits to how the Google Books N-Gram Corpus (Michel et al. 2011) can be used to make inferences about past linguistic and cultural trends. As they show, this corpus lacks appropriate balancing due to an overrepresentation of scientific, academic texts, starting around 1900. This skews the results, making phrases common to academic language appear to increase in frequency when, in fact, the observed increase can be attributed to the change in the types of texts represented in the corpus. Such issues are particularly prevalent to evolutionary-linguistic approaches, as many corpora (and historical corpora necessarily so) are corpora of written language, and the key interest for many evolutionary linguists is face-to-face interactions, 'the core ecology for language use' (Levinson and Holler 2014).

From the definitional remarks above, we have already gathered some important principles of corpus data collection. Corpus compilation, however, is only the first step of corpus-linguistic practice, and it is one that can often be skipped as there are numerous existing, pre-compiled corpora that are suitable for addressing many different research questions. The subsequent steps consist in the exploitation and analysis of corpora. By exploitation—a term fairly common in corpus linguistics (see e.g. McEnery and Wilson 2001)—we mean querying the corpus for the items under investigation and conducting simple quantitative analyses, often aided by specific corpus query tools. But this is usually complemented by more complex analyses, which can be quantitative, qualitative, or—in most cases—a combination of both. Typically, a key step of corpus analyses is the annotation of data, i.e. enriching the data with additional (linguistic or other) information (Gries and Berez 2017; Ide and Pustejovsky 2017). In many cases,

9780192886491_Book.indb 114 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:29

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

corpora come pre-annotated with coding for parts-of-speech (POS) and sometimes lemmas, i.e. the uninflected base forms of words. Such annotations are usually added automatically with the help of NLP tools, so-called taggers. But many annotations also require manual work, especially if complex semantic variables are concerned. For instance, animacy has been shown to be an influential variable for many grammatical phenomena (Yamamoto 1999), and deciding whether a noun or even pronoun refers to a human, an animal, an object, or an abstract concept requires complex in-context decisions (e.g. Zaenen et al. 2004). The annotated datasets can then be analysed using quantitative methods and statistical modelling.

But there are also more explorative approaches that exclusively draw on quantitative methods (see Hills and Miani forthcoming, for a review). For instance, distributional-semantic approaches have become quite popular in corpus linguistics (Boleda 2020). In this approach, the semantic similarity or dissimilarity of linguistic expressions is gauged from the contexts in which they appear. Such an approach can also be used to detect changes over time, e.g. regarding the collocational preferences of a specific construction (Perek 2016) or for drawing conclusions about how a specific word is used in context, e.g. literally vs. figuratively (Schlechtweg et al. 2017). In the following sections, we will discuss how corpus linguistics can inform theories of language evolution, and we will take a closer look at some examples of previous research in which corpus-linguistic approaches have been used to test evolutionary-linguistic hypotheses.

6.3 What can corpus linguistics tell us about language evolution?

There are a number of different ways in which corpus linguistics can inform language evolution research. One arguably under utilized area is testing evolutionary hypotheses about language evolution using corpora of contemporary language use. Szala et al. (2022), for example, investigated Dunbar's (1996) proposal of a socially affiliative adaptive function of language using SpokesPL (Pęzik 2014), a 2.6 million-word corpus of casual, informal conversations in Polish. Dunbar's hypothesis is situated in the context of his 'social brain hypothesis' (Dunbar 2009). Dunbar (1996, 2009) proposed that 'gossiping' allows people to increase cooperation and trust, and thereby increases group cohesion. This enables humans to form larger and more cooperative groups than seen in non-human primates. While Dunbar et al. (1997) provided some evidence for the prevalence of talk about social topics in conversations on a very limited dataset, a detailed corpus-based analysis shows the potential to investigate such claims in much more depth. More specifically, Szala et al. (2022) coded fifty-six conversations for one variable of interest, viz. whether the conversation was about 'social topics', defined as talk about persons or person groups, showing that this was the case for about half of the conversations. While this result is only informative to a limited extent, follow-up studies

9780192886491_Book.indb 115 /12_first_proofs/first_pr…/xml_for_typesetting 27-Mar-25 23:02:29

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

could address the correlation of social conversation topics with other variables in more detail, operationalizing Dunbar's idea of group cohesion using measurable parameters such as the choice of address forms, the frequency of first person plural pronouns or word forms, etc.

Another example is Lev-Ari and Peperkamp (2017), who used corpus data to investigate the role of security vs. insecurity in the environment in grammatical alignment and in the adoption of linguistic norms. Alignment, i.e. adjusting to the environment, has been suggested as a key factor in long-term linguistic change. Using a corpus of forty-one episodes of the TV game show *Jeopardy*, Lev-Ari and Peperkamp found that when participants performed poorly in the show, their reliance on more frequent grammatical structures increased. What this means for theories of the evolutionary dynamics of language change is that not only linguistic and cognitive factors influence linguistic change, but also factors such as success in a particular environment. One particularly interesting aspect of Lev-Ari and Peperkamp's (2017) study is that they also conducted an experimental study in which participants played a game together. They replicated the result from the corpus study that success or lack thereof was related to linguistic alignment. This indicates another fruitful approach to studying the dynamics of language, by combining corpus analyses with psycholinguistic experiments (Baker and Egbert 2016; Meyer et al. 2018; Ventura et al. 2022).

Another way in which corpus research can inform language evolution research is through the use of historical corpora. Here, two kinds of studies are of special importance: first, although many of these studies make no direct reference to language evolution, corpus studies on grammaticalization and the emergence of structure in language can shed light on the hypothesized process of a gradual transition from protolanguage or early languages to fully-fledged, structured modern languages via the incremental emergence of new grammatical structures out of lexical material (Heine and Kuteva 2007; Bybee 2010; Arbib 2012; Hartmann and Pleyer 2021). That is, grammaticalization and other processes operating in language change, as attested by corpus methods, offer mechanisms that can explain how linguistic structure emerged in the first place from an initial set of lexical form–meaning pairings. In addition, these studies also allow us to address the crucial question of how the individual level, as well as the population level, interact in the emergence of structure. Such structure first emerges in individual interactions and instances of language use, then a) becomes increasingly entrenched in the minds of language users, and b) spreads through communities of practice and becomes conventionalized over time (Schmid 2020; Pleyer 2023). One particularly fruitful approach to investigate this process is by relating changes at the community level to changes on the individual level, as attested in high-resolution corpora that allow for tracking changes in individuals' language use through time, such as the ninety million-word corpus of Early Modern Multiloquent Authors (EMMA; Petré et al. 2019). The EMMA corpus collects texts of '50 carefully selected authors across five generations, mostly taken from the 17th-century London society' (Petré et al. 2019: 83). By relating these two levels, researchers can gain insights into the mechanisms of innovation, propagation, and conventionalization, and the way they interact in language evolution

9780192886491_Book.indb 116 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:29

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

(Petré and Van de Velde 2018). For example, investigating the grammaticalization of the English *BE going to* future (in a large corpus that partly overlaps with EMMA), Petré and Van de Velde (2018) show that from the first to the second generation of writers represented in their corpus, the 'motion' meaning of *go* is lost, and authors start to use the construction with inanimate subjects (cf. *it is going to rain*); subsequently, these initially innovative uses become more and more conventional(ized), leading Petré and Van de Velde (2018) to the conclusion that the grammaticalization process starts out on a non-social level with a reanalysis of the structure, followed by a growing degree of social accommodation that leads to a propagation of the new form in the speech community.

Second, some studies have adopted a culturally evolutionary approach, investigating to what degree change in language can be captured in terms of evolutionary dynamics that also characterize the cultural evolution of other cultural systems, using detection methods from the evolutionary sciences (e.g. Lieberman et al. 2007). One instance of this approach will be discussed as our case study in the next section.

6.4 Case study: Disentangling stochastic drift and selection in language evolution

Many different selection forces have been proposed as potential drivers of language change which interact, and at times compete, with each other. These include, for example, analogy (De Smet and Fischer 2017), economy (Zipf 1949; Martinet 1955), simplicity and expressivity (Kirby et al. 2015; Gabelentz [1892] 2016), ease of processing (Hawkins 2004), information content (Piantadosi et al. 2011), language contact (Matras 2020), sociolinguistic prestige (Labov 2011), and a host of other cultural, pragmatic, and cognitive processes (e.g. Croft 2000; Hopper and Traugott 2003; Hundt et al. 2017; Schmid 2020).

However, evolutionary approaches to language change have increasingly looked to methods from population genetics and cultural evolution to shed light on mechanisms influencing the evolution of language. One particular process that has been investigated is that of 'stochastic drift', that is, random fluctuations in how linguistic forms are copied between language users. In population genetics and cultural evolution, this process has been shown to explain the change of different variants, with random fluctuations over time accumulating to produce substantive changes. For language evolution, this means that the change from one linguistic form to another might simply occur by random chance instead of being driven by a particular selective force. Whereas a number of approaches have mostly relied on mathematical models and simulations to investigate this process (e.g. Reali and Griffiths 2010; Blythe 2012), others have used corpus data of different kinds (e.g. Lieberman et al. 2007; Bentley 2008; Sindi and Dale 2016; Newberry et al. 2017).

9780192886491_Book.indb 117 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:29

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Adopting a method to detect stochastic drift from population genetics—the Frequency Increment Test (FIT)—Newberry et al. (2017) analysed three historical changes in English to test if they were the result of selection or drift. The FIT was originally developed in genetics in order to disentangle the respective influence of genetic drift and natural selection on the frequency variation of alleles in a population. It tests the neutral null hypothesis that changes in allele frequency are caused only by genetic drift. Newberry et al. (2017) applied this method to historical corpora. Specifically, using historical corpus data, they investigated signals of drift vs. selection in the regularization of English past tense verbs (e.g. *sneaked* vs. *snuck*), the introduction of the periphrastic *do* in the Early Modern English period (*You say not* > *You do not say*), and an instance of Jespersen's cycle of negation that occurred in Middle English (*Ic ne sedge* > *I ne seye not* > *I say not*).[2] In the first two cases, Newberry et al. (2017) found evidence of drift in some contexts, whereas, in others, the null hypothesis of neutral evolution could be rejected, indicating selection. In the third case, that of negation, stochastic drift could be rejected, in line with research in historical linguistics that suggests that this change is due to a selective process in which language users overuse negation forms which, in turn, increasingly lose emphatic force (van der Auwera 2009). Overall, the authors see their results as revealing 'an underappreciated role for stochasticity in language evolution' (Newberry et al. 2017: 223).

The result that has received most attention—in the sense that it was scrutinized in subsequent research to which we return below—is Newberry et al.'s (2017) investigation of changes in past tense forms. Newberry et al. used data from the Corpus of Historical American English (COHA), a 400 million-word corpus covering texts between 1810 and 2009 (Davies 2012).

They then calculated the relative frequencies of past tense forms of thirty-six verbs that at some point in their development exhibited different past tense formation patterns (regular, e.g. *lighted*, *quitted*, *throwed*, *awaked*, vs. irregular, e.g. *lit*, *quit*, *threw*, *awoke*). These were then placed into 'bins' (i.e. sections) of variable lengths, and the FIT was used to analyse the resulting historical patterns of individual verbs, with each bin essentially representing a 'generation' in analogy with population genetics. However, the temporal granularity of Newberry et al.'s bins differed considerably, as they opted for a variable binning strategy so that each bin has roughly the same number of tokens. Thus, their bin lengths ranged from fifty-three years to fifteen years (see Karjus et al. 2020). Karjus et al. (2020) replicated Newberry et al.'s study, but with different binning strategies. Whereas the number of verbs exhibiting signals of selection was similar across binning strategies, the specific verbs that exhibited these signals differed with binning strategy. Four verbs (*light*, *smell*, *sneak*, *wake*) exhibited signals

[2] Jespersen's cycle refers to the phenomenon that in many languages, the pre-verbal negation element tends to lose substance and is therefore strengthened by the addition of a post-verbal reinforcement, which can become obligatory and even end up as the sole expression of negation, e.g. Early French *Jeo ne dis* 'I don't say' > Middle French *Je ne dis pas* > contemporary informal French *je dis pas* (Larrivée 2011; Jespersen 1917).

9780192886491_Book.indb 118 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:29

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

of selection in most binning choices, but the others differed depending on binning strategy. Moreover, there were no verbs that exhibited unambiguous signals of selection, and all were sensitive to binning to some degree. Karjus et al. (2020) conclude that whereas the FIT method proposed by Newberry et al. works well to detect general trends (i.e. the general trend that more verbs in their dataset seem to change due to stochastic drift than due to selection), it is less well suited to test whether or not a particular item has been subject to selection during language evolution. Overall, they argue that when adopting tests from different domains, such as population genetics, the specific properties of the analysed data (in this case historical corpus data) have to be taken into account. This means that statistical methods such as the FIT can be valuable for certain aims, but are less suited for others. One that is often of special interest to historical linguists is not only to acknowledge that a particular linguistic change has been due to some selection pressure, but to find out which of the many proposed selective drivers of language evolution operate in a particular change (cf. Karjus et al. 2020). For the future, this calls for a closer integration of statistical methods from population genetics and cultural evolution on the one hand, and historical corpus research on specific driving forces on the other.

One approach to overcome the challenges the FIT faces in detecting stochastic drift and selection has been proposed by Karsdorp et al. (2020). They address the key challenge posed by the incomplete and sometimes sparse data in historical corpora, which necessitates the binning of data, which in turn, as Karjus et al. (2020) have shown, influences the results of approaches such as the FIT.

Karsdorp et al. (2020) reformulate the issue of how to detect drift and selection in language evolution as a binary time series classification (TSC) problem. That is, they take the problem of binning as their starting point. To solve it, they train a type of deep neural network architecture that learns to classify whether a particular time series is an instance of selection or drift. They then apply this machine learning mechanism to the same thirty-six verbs used in Newberry et al. (2017) and Karjus et al. (2020) and show that this approach is much less influenced by different binning strategies and is overall able to learn to classify linguistic changes (or more precisely: time series) as instances of either selection or drift in a consistent, efficient, and accurate manner. So, regardless of binning strategy, twenty-four verbs are consistently classified as examples of drift. Moreover, the neural network also classified the same words as the FIT as being subject to selection. However, in addition, it also relatively consistently classified a number of verbs as examples of selection which were not identified by FIT, which, in general, has trouble detecting low selection pressure.

Overall, the role of stochastic drift in language evolution is certainly an aspect that deserves serious consideration, but it should be combined with research on the factors that influence selection in those cases where there is evidence for it. For example, Guerrero Montero et al. (2023) show how computational methods can detect selective forces such as phonological simplicity and socially motivated language change in corpus data. As the studies by Karjus et al. (2020) and Karsdorp et al. (2020) show, this is very much a developing field. One promising area, to gain more insight into the

9780192886491_Book.indb 119 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:29

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

role of drift, is that of complementing corpus-based approaches such as these with experimental studies, such as artificial language learning studies (see Tal et al., Chapter 2, this volume). In contrast to corpus studies, these have the advantage that they allow tracking the entire trajectory of a language and to control for different selecting factors (Galantucci et al. 2012; Ventura et al. 2022; see also Müller and Raviv, Chapter 3, this volume; Derex and Brand, Chapter 14, this volume). This is also the reason why experiments in the cultural evolution of language have been called 'the linguists' drosophila' (Roberts 2017), referring to the fly species that has been used widely in genetic research, and a continuing integration of corpus-based and experimental methods seems highly promising.

6.5 Conclusion

Summing up, corpus-based approaches are a promising and arguably underused method for testing hypotheses about the evolutionary dynamics of language. In this chapter, we have particularly focused on two areas in which corpus studies can prove informative: on the one hand, they can help us assess the respective role of selection and random drift in language change, which also has important implications for the question of how linguistic structure emerged in the first place. On the other hand, they can help test hypotheses about specific factors that drive linguistic change apart from mere random drift, including pragmatic, social, and interactional factors, as in the *Jeopardy* example discussed above. In addition, they also have potential to shed light on the question of how innovations first arise, and then spread through and become conventionalized in a community, a process that is central to any evolutionary account of language.

The examples we have discussed in the present chapter of course cannot give an exhaustive overview of the wealth of corpus-linguistic approaches that have implications for theories of language dynamics and language evolution. While we have focused on pragmatic interaction as well as broad-scope historical approaches, corpus-based approaches could also be used to track the evolutionary dynamics involved in the development of phonological systems (Ritt 2004) or writing systems (Berg and Aronoff 2017; Kelly et al. 2021), which in turn can inform theories of how linguistic structure emerges and is constantly re organized both in individual speakers' minds as well as on the level of an entire population.

Given that corpus data are limited to extant, fully-fledged languages, corpus studies have to be complemented with other approaches in order to put evolutionary-linguistic theories and hypotheses to the test. But as an approach that allows for investigating language dynamics 'in the wild', rather than in controlled laboratory settings, corpus linguistics provides an important puzzle piece for understanding the evolutionary dynamics of language.

9780192886491_Book.indb 120 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:29

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

References

Arbib, M. A. (2012). *How the Brain Got Language: The Mirror Systems Hypothesis*. Oxford: Oxford University Press.

Armstrong, D. F. and Wilcox, S. (2007). *The Gestural Origin of Language*. Oxford: Oxford University Press.

Baker, P. and Egbert, J. (eds.). (2016). *Triangulating Methodological Approaches in Corpus-Linguistic Research*. New York: Routledge.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). 'The WaCky Wide Web: A collection of very large linguistically processed web-crawled corpora', *Language Resources and Evaluation* 43(3): 209–226.

Behrens, H. (eds.). (2008). *Corpora in Language Acquisition Research: History, Methods, Perspectives*. Amsterdam, Philadelphia: John Benjamins.

Bentley, R. A. (2008). 'Random drift versus selection in academic vocabulary: An evolutionary analysis of published keywords', *PLOS ONE. Public Library of Science* 3(8): e3057, https://doi.org/10.1371/journal.pone.0003057.

Berg, K. and Aronoff, M. (2017). 'Self-organization in the spelling of English suffixes: The emergence of culture out of anarchy', *Language* 93(1): 37–64.

Bergmann, T. and Dale, R. (2016). 'A scientometric analysis of Evolang: Intersections and authorships', in S. G. Roberts, C. Cuskley, L. McCrohon, et al. (eds.), *The Evolution of Language. Proceedings of the 11th International Conference*, 79–87, http://evolang.org/neworleans/papers.

BNC Consortium (2007). The British National Corpus, XML Edition, Oxford Text Archive, http://hdl.handle.net/20.500.12024/2554 (last checked 02/03/2023).

Blythe, R. A. (2012). 'Neutral evolution: A null model for language dynamics', *Advances in Complex Systems* World Scientific Publishing Co. 15(03n04): 1150015, https://doi.org/10.1142/S0219525911003414.

Boleda, G. (2020). 'Distributional semantics and linguistic theory', *Annual Review of Linguistics* 6(1): 213–234, https://doi.org/10.1146/annurev-linguistics-011619-030303.

Bybee, J. L. (2010). *Language, Usage and Cognition*. Cambridge: Cambridge University Press.

Croft, W. (2000). *Explaining Language Change: An Evolutionary Approach*. Harlow: Longman.

Davies, M. (2012). 'Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English', *Corpora* 7(2): 121–157, https://doi.org/10.3366/cor.2012.0024.

De Smet, H. and Fischer, O. (2017). 'The role of analogy in language change: Supporting constructions', in M. Hundt, S. Mollin, and S. E. Pfenninger (eds.), *The Changing English Language: Psycholinguistic Perspectives* (Studies in English Language), 240–268. Cambridge: Cambridge University Press, https://doi.org/10.1017/9781316091746.011.

Dunbar, R. I. M. (1996). *Grooming, Gossip, and the Evolution of Language*. Cambridge: Cambridge University Press.

Dunbar, R. I. M. (2009). 'The social brain hypothesis and its implications for social evolution', *Annals of Human Biology* 36(5): 562–572, https://doi.org/10.1080/03014460902960289.

Dunbar, R. I. M., Marriott, A., and Duncan, N. D. C. (1997). 'Human conversational behavior', *Human Nature* 8(3): 231–246, https://doi.org/10.1007/BF02912493.

Galantucci, B., Garrod, S., and Roberts, G. (2012). 'Experimental semiotics', *Language and Linguistics Compass* 6(8): 477–493, https://doi.org/10.1002/lnc3.351.

9780192886491_Book.indb 121 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:29

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Gries, S. Th. and Berez, A. L. (2017). 'Linguistic annotation in/for corpus linguistics', in N. Ide and J. Pustejovsky (eds.), *Handbook of Linguistic Annotation*, 379–409. Dordrecht: Springer.

Guerrero Montero, J., Karjus, A., Smith, K., and Blythe, R. A. (2023). 'Reliable detection and quantification of selective forces in language change', *Corpus Linguistics and Linguistic Theory*, https://doi.org/10.1515/cllt-2023-0064.

Gut, U. (2020). 'Spoken corpora', in M. Paquot and S. Th. Gries (eds.), *A Practical Handbook of Corpus Linguistics*, 235–256. Cham: Springer International Publishing, https://doi.org/10.1007/978-3-030-46216-1_11.

Hartmann, S. and Pleyer, M. (2021). 'Constructing a protolanguage: Reconstructing prehistoric languages in a usage-based construction grammar framework', *Philosophical Transactions of the Royal Society B: Biological Sciences* 376(1824): rstb.2020.0200, 20200200, https://doi.org/10.1098/rstb.2020.0200.

Hartmann, S., Wacewicz, S., Ravignani, A., et al. (2024). 'Delineating the field of language evolution research: A quantitative analysis of peer-review patterns at the Joint Conference on Language Evolution (JCoLE 2022)'. *Interaction Studies* 25(1): 100–117, https://doi.org/10.1075/is.00024.har.

Hawkins, J. A. (2004). *Efficiency and Complexity in Grammars*. Oxford: Oxford University Press.

Heine, B. and Kuteva, T. (2007). *The Genesis of Grammar: A Reconstruction*. Oxford: Oxford University Press.

Hills, T. and Miani, A. forthcoming. 'A short primer on historical natural language processing', in G. Progrebna and T. Hills (eds.), *Handbook of Behavioural Data Science*. Cambridge: Cambridge University Press.

Hodge, G. and Goico, S. A. (2022). 'Natural and elicited: Sign language corpus linguistics and linguistic ethnography as complementary methodologies', *Journal of Sociolinguistics* 26(1): 126–136, https://doi.org/10.1111/josl.12523.

Hopper, P. J. and Traugott, E. C. (2003). *Grammaticalization*. 2nd edn. Cambridge: Cambridge University Press.

Hundt, M., Mollin, S., and Pfenninger, S. E. (eds.). (2017). 'The changing English language', *The Changing English Language: Psycholinguistic Perspectives* (Studies in English Language). Cambridge: Cambridge University Press.

Hunston, S. (2008). 'Collection strategies and design decisions', in A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics: An International Handbook (HSK 29.1)*, 154–168. Berlin, New York: Walter de Gruyter.

Ide, N. and Pustejovsky, J. (eds.). (2017). *Handbook of Linguistic Annotation*. Dordrecht: Springer.

Jakubíček, M., Kilgarriff, A., Kovář, V., et al. (2013). 'The TenTen corpus family', *7th International Corpus Linguistics Conference*. Lancaster: University of Lancaster, http://ucrel.lancs.ac.uk/cl2013/doc/CL2013-ABSTRACT-BOOK.pdf. (20 November 2019).

Jespersen, O. (1917). *Negation in English and Other Languages*. Copenhaguen: Høst.

Karjus, A., Blythe, R. A., Kirby, S., and Smith, K. (2020). 'Challenges in detecting evolutionary forces in language change using diachronic corpora', *Glossa: A Journal of General Linguistics. Open Library of Humanities* 5(1), https://doi.org/10.5334/gjgl.909.

Karsdorp, F., Manjavacas, E., Fonteyn, L., and Kestemont, M. (2020). 'Classifying evolutionary forces in language change using neural networks', *Evolutionary Human Sciences* 2: e50, https://doi.org/10.1017/ehs.2020.52.

9780192886491_Book.indb 122 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:29

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Kelly, P., Winters, J., Miton, H., and Morin, O. (2021). 'The predictable evolution of letter shapes: An emergent script of West Africa recapitulates historical change in writing systems', *Current Anthropology* 62(6): 669–691, https://doi.org/10.1086/717779.

Kilgarriff, A. (2007). 'Googleology is bad science', *Computational Linguistics* 33(1): 147–151, https://doi.org/10.1162/coli.2007.33.1.147.

Kirby, S., Tamariz, M., Cornish, H., and Smith, K. (2015). 'Compression and communication in the cultural evolution of linguistic structure', *Cognition* 141: 87–102.

Kübler, S. and Zinsmeister, H. (2015). *Corpus Linguistics and Linguistically Annotated Corpora*. London: Bloomsbury.

Labov, W. (1972). *Sociolinguistic Patterns (Conduct and Communication 4)*. Philadelphia: University of Pennsylvania Press.

Labov, W. (2011). *Principles of Linguistic Change, Volume 3: Cognitive and Cultural Factors*. New York: Wiley-Blackwell.

Larrivée, P. (2011). 'Is there a Jespersen cycle?', in P. Larrivée and R. Ingham (eds.), *The Evolution of Negation*, 1–22. Berlin, Boston: De Gruyter, https://doi.org/10.1515/9783110238617.1.

Lemnitzer, L. and Zinsmeister, H. (2015). *Korpuslinguistik. Eine Einführung*. 3rd edn. Tübingen: Narr.

Lev-Ari, S. and Peperkamp, S. (2017). 'Language for $200: Success in the environment influences grammatical alignment', *Journal of Language Evolution* 2(2): 177–187, https://doi.org/10.1093/jole/lzw012.

Levinson, S. C. and Holler, J. (2014). 'The origin of human multi-modal communication', *Philosophical Transactions of the Royal Society B: Biological Sciences* 369(1651): 20130302, https://doi.org/10.1098/rstb.2013.0302.

Lieberman, E., Michel, J.-B., Jackson, J., et al. (2007). 'Quantifying the evolutionary dynamics of language', *Nature* 449(7163): 713–716, https://doi.org/10.1038/nature06137.

Lüdeling, A. and Kytö, M. (eds.). (2008). *Corpus Linguistics: An International Handbook* (HSK 29.1). Berlin, New York: Walter de Gruyter.

Martinet, A. (1955). *Économie des changements phonétiques. Traité de phonologie diachronique*. Bern: Francke.

Matras, Y. (2020). *Language Contact*. 2nd edn. Cambridge: Cambridge University Press.

McEnery, T. and Wilson, A. (2001). *Corpus Linguistics: An Introduction*. 2nd edn. Edinburgh: Edinburgh University Press.

Mendívil-Giró, J.-L. (2019). 'Did language evolve through language change? On language change, language evolution and grammaticalization theory', *Glossa: A Journal of General Linguistics* 4(1): 124, https://doi.org/10.5334/gjgl.895.

Meyer, A. S., Alday, P. M., Decuyper, C., and Knudsen, B. (2018). 'Working together: Contributions of corpus analyses and experimental psycholinguistics to understanding conversation', *Frontiers in Psychology* 9, https://www.frontiersin.org/articles/10.3389/fpsyg.2018.00525. (20 March 2023).

Michel, J.-B., Shen, Y. K., Aiden, A. P., et al. (2011). 'Quantitative analysis of culture using millions of digitized books', *Science* 331(6014): 176–182, https://doi.org/10.1126/science.1199644.

Newberry, M. G., Ahern, C. A., Clark, R., and Plotkin, J. B. (2017). 'Detecting evolutionary forces in language change', *Nature* 551(7679): 223–226, https://doi.org/10.1038/nature24455.

Paquot, M. and Gries, S. Th. (eds.). (2020). *A Practical Handbook of Corpus Linguistics*. Cham: Springer International Publishing.

9780192886491_Book.indb 123 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:29

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Pechenik, E. A., Danforth, C. M., and Dodds, P. S. (2015). 'Characterizing the Google Books Corpus: Strong limits to inferences of socio-cultural and linguistic evolution', *PLoS ONE* 10(10): e0137041, https://doi.org/10.1371/ journal.pone.0137041.

Perek, F. (2016). 'Using distributional semantics to study syntactic productivity in diachrony: A case study', *Linguistics* 54(1): 149–188.

Perlman, M. (2017). 'Debunking two myths against vocal origins of language: Language is iconic and multimodal to the core', *Interaction Studies* 18(3): 376–401, https://doi.org/10.1075/is.18.3.05per.

Petré, P. and Van de Velde, F. (2018). 'The real-time dynamics of the individual and the community in grammaticalization', *Language* 94(4): 867–901.

Petré, P., Anthonissen, L., Budts, S., et al. (2019). 'Early Modern Multiloquent Authors (EMMA): Designing a large-scale corpus of individuals' languages', *ICAME Journal* 43(1): 83–122.

Pęzik, P. (2014). 'Spokes: A search and exploration service for conversational corpus data', in Jan Odijk (ed.), *Selected Papers from the CLARIN 2014 Conference*, 99–109. Linköping: Linköping University Electronic Press, https://ep.liu.se/ecp/116/ecp15116.pdf (accessed 16 December 2024).

Piantadosi, S. T., Tily, H., and Gibson, E. (2011). 'Word lengths are optimized for efficient communication', *Proceedings of the National Academy of Sciences* 108(9): 3526–3529.

Pleyer, M. (2023). 'The role of interactional and cognitive mechanisms in the evolution of (proto)language(s)', *Lingua* 282: 103458, https://doi.org/10.1016/j.lingua.2022.103458.

Pleyer, M. and Hartmann, S. (2024). *Cognitive Linguistics and Language Evolution*. Cambridge: Cambridge University Press, https://doi.org/10.1017/9781009385022.

Reali, F. and Griffiths, T. L. (2010). 'Words as alleles: Connecting language evolution with Bayesian learners to models of genetic drift', *Proceedings of the Royal Society B: Biological Sciences* 277(1680): 429–436, https://doi.org/10.1098/rspb.2009.1513.

Ritt, N. (2004). *Selfish Sounds and Linguistic Evolution: A Darwinian Approach to Language Change*. Cambridge: Cambridge University Press.

Roberts, G. (2017). 'The linguist's Drosophila: Experiments in language change', *Linguistics Vanguard* 3(1): 20160086, https://doi.org/10.1515/lingvan-2016-0086.

Schäfer, R. and Bildhauer, F. (2012). 'Building large corpora from the web using a new efficient tool chain', in N. Calzolari, K. Choukri, T. Declerck, et al. (eds.), *Proceedings of LREC 2012*, 486–493.

Schlechtweg, D., Eckmann, S., Santus, E., et al. (2017). 'German in flux: Detecting metaphoric change via word entropy', in *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, 354–367. Vancouver: Association for Computational Linguistics, https://doi.org/10.18653/v1/K17-1036.

Schmid, H.-J. (2020). *The Dynamics of the Linguistic System: Usage, Conventionalization, and Entrenchment*. Oxford: Oxford University Press.

Sindi, S. S. and Dale, R. (2016). 'Culturomics as a data playground for tests of selection: Mathematical approaches to detecting selection in word use', *Journal of Theoretical Biology* 405: 140–149, https://doi.org/10.1016/j.jtbi.2015.12.012.

Stefanowitsch, A. (2020). *Corpus Linguistics: A Guide to the Methodology*. Berlin: Language Science Press.

Szala, A., Placiński, M., Poniewierska, A., et al. (2022). 'How much language use is actually on social topics?', in A. Ravignani, R. Asano, D. Valente, et al. (eds.), *The Evolution of*

9780192886491_Book.indb 124 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:30

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Language: Proceedings of the Joint Conference on Language Evolution (JCoLE), 705–707. Joint Conference on Language Evolution (JCoLE), https://doi.org/10.17617/2.3398549.

Szala, A., Wacewicz, S., Placiński, M., et al. (2025). 'How much conversation content is actually social? Human conversational behaviour revisited', *Language and Cognition* 17, https://doi.org/10.1017/langcog.2024.54.

Taylor, C. (2008). 'What is corpus linguistics? What the data says', *ICAME Journal* 32: 179–200.

van der Auwera, J. (2009). 'The Jespersen cycles', in E. van Gelderen (ed.), *Cyclical Change*, 35–71. Amsterdam, Philadelphia: John Benjamins.

Ventura, R., Plotkin, J. B., and Roberts, G. (2022). 'Drift as a driver of language change: An artificial language experiment', *Cognitive Science* 46(9): e13197, https://doi.org/10.1111/cogs.13197.

von der Gabelentz, G. (2016). *Die Sprachwissenschaft. Ihre Aufgaben, Methoden und bisherigen Ergebnisse, ed. James McElvenny and Manfred Ringmacher*. Berlin: Language Science Press.

Wacewicz, S., Sibierska, M., Placiński, M., et al. (2022). 'The scientometric landscape of Evolang: A comprehensive database of the Evolang conference'. *Journal of Language Evolution* 7(2): 246–258, https://doi.org/10.1093/jole/lzad003.

Wacewicz, S., Sibierska, M., Placiński, M., et al. (2023). 'The scientometric landscape of Evolang: A comprehensive database of the Evolang conference', *Journal of Language Evolution*. doi 10.1093/jole/lzad003.

Yamamoto, M. (1999). *Animacy and Reference: A Cognitive Approach to Corpus Linguistics* (Studies in Language Companion Series 46). Amsterdam, Philadelphia: John Benjamins.

Zaenen, A., Carletta, J., Garretson, G., et al. (2004). 'Animacy encoding in English: Why and how', in B. Webber and D. Byron (eds.), *DiscAnnotation '04*, 118–125. Stroudsburg, PA: Association for Computational Linguistics.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Oxford: Addison-Wesley Press.

9780192886491_Book.indb 125 /12_first_proofs/first_pr[illegible]/xml_for_typesetting 27-Mar-25 23:02:30

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

9780192886491_Book.indb 126 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:30

CHAPTER 7

ADAPTATION

CALEB EVERETT

7.1 Background: Cultural evolution, language evolution

Human cultures are adapted for adaptation. Our circum-ambulation of the globe, largely completed dozens of millennia ago, was enabled by this characteristic. Humans have, of course, adapted to diverse environments partially through genetic adaptations as well. Larger, thicker bodies are characteristic of colder climates, possibly because they reduce the ratio of surface area to body volume and inhibit heat loss via radiation. This point, often referred to as Bergmann's rule, holds for other mammals as well (Meiri and Dayan 2003). Lighter skin associates with higher latitudes, likely due to its facilitated synthesis of Vitamin D (Hanel and Carlberg 2020). Our faces and nasal cavities have adapted to better humidify inhaled air in very cold environments (Evteev et al. 2014). These and other genetic adaptations have helped promote the success of our species across environments (see e.g. Fan et al. 2016). Yet, cross-population genetic variation has played a relatively minor role in our species' adaptation to diverse environments. This is evidenced by the very limited genetic diversity of *sapiens*, when contrasted to other primates, and the relatively shallow time depth of the global spread of populations from Africa (see e.g. Prugnolle et al. 2005). Further, many of the physiological differences across human populations may be due to genetic drift and have no clear adaptive features. Instead, the dominant domain of human adaptation, enabling our success in so many distinct environments, is culture. As research on cultural evolution has demonstrated over the past few decades, many of the adaptive behaviours of humans are the product of selective processes at the level of cultures rather than individuals. Furthermore, cultures may now place distinct pressures on the selection of genes; stronger even than the pressures of particular physical environments (Richerson and Boyd 2008).

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

As in the case of genetic changes, many instances of cultural change do not represent adaptations but might be considered more analogous to genetic drift. Simultaneously, though, cultures are continually producing behavioural variants that might be selected because they confer some advantage in particular environments. Many such adaptive variants are selected non-consciously and without the members of a culture necessarily being aware of the motivations for the behaviour's success. Instead, prestige may be conferred to individuals exhibiting the successful variant, and this prestige might lead to the emulation of the relevant behaviour even if individuals do not know why exactly the behaviour is successful. Many examples could be adduced, for instance, the elaborate processing of manioc prior to ingestion, despite the fact that many who use such processing techniques are unaware of the specific ways in which they reduce hydrocyanic acid (Henrich 2015). To cite another example of many, the ingestion of cod-fish oil, which increases vitamin D consumption and reduces the risk of osteomalacia, has been practised in certain high-latitude cultures for centuries without an awareness of the existence, let alone the benefits, of vitamin D (Ringe 1998). Such behaviours may lead to the 'success' of particular cultures, as the advantages of the behaviours disperse through the members of the cultures in question. Other behavioural variants may yield more modest benefits to individuals, simply making their life easier and reducing effort for everyday tasks.

It is unclear why language would not share the adaptive characteristics of cultural evolution and why, more specifically, languages would not exhibit externally influenced adaptive characteristics that reduce the effort required for communication (to some, language change is an example of cultural evolution but only exhibits the drift-like characteristics of cultural evolution, not the adaptive ones). This point could apply to varied linguistic phenomena, from lexemes to phonemes. With respect to the latter, some researchers now suggest that the default hypothesis should be that human sound systems are shaped in subtle ways by environmental factors. Wang et al. (2023: 1) state that 'it is almost self-evident that the phonetic system must be affected by the environment, since language communication predominantly relies on sound as the medium, and both the production and transmission of speech sounds are susceptible to external factors'. Nevertheless, while it is generally acknowledged in linguistics that the pressure for reduction of effort of communication does influence how languages evolve, until recently, it was not considered whether this pressure yields distinct diachronic trajectories of sound systems owing to adaptive-like pressures that vary cross-culturally. Below, I discuss some recent evidence suggesting that the sounds used in the world's languages, like other facets of human cultures, exhibit such adaptive characteristics. I also illustrate the general method that is being used to tackle this issue, while drawing attention to some of the limitations of the approach.

Before delving into the method and some of the proposed adaptive characteristics exhibited by the world's sound systems, it is worth underscoring that lexical adaptation is clearly evident across cultures. One oft-cited example is that distinct words for 'snow' and 'ice' are more likely to exist in cold climates, while these concepts are much more likely to be encoded with one lexeme in warm-weather climates (Regier et al. 2016). Less

9780192886491_Book.indb 128 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:30

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

obvious lexical adaptation exists, which may interact with other cultural processes. For example, Gibson et al. (2017) describe the ways in which colour lexicons have adapted to the need for more refined communication regarding brightly coloured foregrounded objects in a post-industrialized world. Spatial terms are used in ways that are environmentally adaptive, as evidenced by experimental research carried out in virtual reality contexts (Nölle et al. 2020). Beyond such lexical adaptation, it has also been suggested that languages adapt to external pressures at the morphological level. Lupyan and Dale (2010) posited that languages spoken by large populations develop reduced morphological complexity to facilitate acquisition by adult learners. Recent evidence calls this claim into doubt, however (Shcherbakova et al. 2023). Yet other work has suggested, via experimental methods, that languages adapt to distinct communicative pressures placed upon them by larger populations (Raviv et al. 2019). Despite ongoing debates, these and other findings have led to the increased realization that languages adapt to external pressures in ways that have gone largely unnoticed (Bentz et al. 2018).

Recent studies have proposed that the sound systems of languages adapt to anatomical variations across populations, or to environmental factors that interact with human physiology. In most of the proposed adaptations, the underlying uncontroversial mechanism is a bias towards sound types that require less effort to produce (other hypotheses suggest perceptual biases are at work in environment–phonetic interactions—see e.g. Maddieson and Benedict 2023 and Wang et al. 2023). However, some of the proposals relying on this underlying uncontroversial mechanism are themselves controversial. Two potential cases of adaptation to cross-population differences in anatomy are offered in Moisik and Dediu (2017) and Dediu et al. (2019). In the former study, the authors suggest that the smoother palate in some populations, one characterized by a less pronounced alveolar ridge, facilitates the production of click consonants that are so common among Sub-Saharan languages. In the latter study, the authors suggest that distinct palatal configurations bias languages to rely on certain vowel types, given sufficient time (see also the discussion in Dediu et al. 2017). Relatedly, Blasi et al. (2019) suggest that populations with overbite and overjet dental features are more likely to rely on labiodental consonants, i.e. that languages adapt to differences in dentition across populations. This hypothesis is one of two discussed in the following section in greater detail, to illustrate the general methodology evident in studies on the potential adaptive characteristics of sound systems. Given that such bite characteristics are more typical in cultures with softer diets, this hypothesis points to an adaptive characteristic of speech that relates to subsistence type and environmental characteristics that favour certain subsistence strategies (Everett and Chen 2021).

In a similar vein, some studies have suggested that certain articulatory gestures relying on the vocal cords may be inhibited in specific environments given the effects of ambient air characteristics on the operation of the vocal cords, effects which have been demonstrated in much experimental work in laryngology (Alves et al. 2019). This includes work on complex tonality and the apparent inhibition of its development by extreme aridity (Everett et al. 2015; Liang et al. 2023). It also includes the suggestion that ejective consonants may require reduced articulatory effort in regions with reduced

9780192886491_Book.indb 129 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:30

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

ambient air pressure (Everett 2013, though this latter suggestion is based on simple theoretical modelling only, not on experimental work). Potential issues with these related hypotheses are raised, *inter alia*, in De Boer (2016) and Moran and Urban (2021). An interrelated hypothesis is presented in Everett (2017), in which I suggest that languages might generally rely less on vowels in very arid environments. The broader hypothesis at work in studies like Liang et al. (2023) and Everett (2017) is that very dry ambient air can impact the operation of the vocal cords to the extent that the evolution of sound systems is affected. The effect of very dry ambient air on the vocal cords, even over short durations, is well established in experimental laryngology (Sundarrajan et al. 2017). The claim that this effect is sufficient to impact the use of vowels and tonality, subsumed here under one hypothesis given that these are interrelated claims (see Everett 2021), are less well established. This hypothesis, along with the labiodentals hypothesis mentioned above, is discussed in greater detail in the next two sections. Both hypotheses are ultimately based on the fact that some sound changes owe themselves to the reduction of articulation effort, other factors being equal.

7.2 Outline of method

The method used to explore potential phonetic/phonological adaptations to external factors could be described via a series of four steps. The steps can be encapsulated, i–iv, as follows:

i) Begin with a known finding on human anatomy or a finding on the potential interaction of human anatomy/physiology and the physical environment, from a discipline outside linguistics. For instance, in the case of Everett et al. (2015) and Everett (2017), the key findings are taken from the field of laryngology. The relevant experimental findings suggest that extremely arid air impacts the functioning of the vocal cords by reducing their viscosity and elasticity (e.g. Leydon et al. 2009; Sundarrajan et al. 2017; Alves et al. 2019). In the case of Blasi et al. (2019) and Everett and Chen (2021), the non-linguistic findings relate to dental anatomy in a broad sense, i.e. that bite types differ across populations and across life stages. These differences also relate to other aspects of human behaviour, namely variations in diet across populations: people with softer diets more typically develop the dental features of overbite and overjet, when contrasted to people with rougher diets such as hunter-gatherers. For instance, in the USA the average adult has an overbite of 3 mm, and an overjet of roughly the same amount (Brunelle et al. 1996). In this case, the findings are not anatomical alone, as they relate to other aspects of culture, namely subsistence strategy and utensil usage. Furthermore, these aspects of cultures are impacted indirectly by geographic factors, since some regions are more likely to yield higher agricultural reliance (Vilela et al. 2020). The cross-population differences in bite types are relevant to

9780192886491_Book.indb 130 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:30

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

speech and, potentially, to the ways languages adapt to external pressures. The same could be said of the effects of extreme aridity on the vocal cords: it has been demonstrated that very dry ambient air impacts the operation of the vocal cords and that perceived effort of vocal cord vibration may increase even after brief exposure of the vocal cords to very dry air (Sivasankar and Fisher 2002). This fact motivates the hypothesis that very dry ambient air impacts the ways that, with time, languages rely on vocal cord vibration.

ii) Given the potential relevance of external factors to speech, uncovered via step (i), specific hypotheses can be generated regarding the long-term development of languages. These hypotheses are fundamentally centred on a mechanism that is well known in linguistics and other disciplines—the bias towards reduction of effort in human behaviour. The specific hypotheses generated at this stage should be consistent with what we know about other relevant phenomena, for instance, speech perception and sound change. Whether or not this consistency exists is open to some debate. For instance, the hypothesis raised by Everett et al. (2015, 2016a) was criticized by some as being inconsistent with some aspects of tone genesis in the literature (e.g. Ladd 2016). In a reply, we suggested that previous research on tone genesis did not consider the possibility of such external factors, implying that the traditional accounts themselves may face limitations while stressing that there are many aspects of tone genesis that are not fully elucidated (Everett et al. 2016b). In general, though, there must be some plausible diachronic mechanism through which the hypothesized effect in question might be realized, and the proposed diachronic effect should not be wholly inconsistent with our understanding of interrelated phenomena like sound change, speech perception, or sociolinguistic variation. This point is discussed in Everett (2021), wherein I offer some ideas on how such adaptive effects could be reified through well-known diachronic processes.

Sticking with the examples being focused on in this discussion, the hypothesis discussed in Blasi et al. (2019), owing ultimately to the work of Hockett (1985), is that the languages of cultures with softer diets should be more likely to use labiodental phonemes. An ancillary hypothesis is that labiodental consonants are a relatively recent innovation, or at least likely to have been much rarer during the Neolithic, prior to the widespread advent of agriculture. In Everett (2017), the specific prediction generated is that the increased effort of vocal cord usage in extremely arid contexts will place pressures on languages in such environments to rely less heavily on vowels. Note that this prediction is unrelated to the number of vowels in a language's phoneme inventor, contra Roberts (2018), but is instead related to the ratio of vowels used in speech. For example, languages with small vowel inventories that simultaneously avoid consonant clusters have high vowel ratios. Relatedly, Everett et al. (2015, 2016a) and Liang et al. (2023) suggest that the increased effort associated with precise phonation in very dry environments may yield pressures against the development of complex tonality. The notion that supralaryngeal air pressure impacts the vocal cords and sound usage more

9780192886491_Book.indb 131 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:30

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

broadly has been established for decades (e.g. Ohala and Ohala 1993), and has been demonstrated clearly in the recent typological literature (Everett 2018). With that in mind, the broad hypothesis underlying such specific claims, namely that ambient air characteristics impact sound usage across languages, is not particularly radical.

iii) The next step in this method is to test the hypothesis in (ii) via correlational tests with relevant databases. These tests consider specific variables relevant to the hypothesis, in contrast to correlational approaches in which databases are first tested for any correlations, with potential causal explanations only generated afterwards (see Roberts and Winters 2013). Given the exploratory nature of the hypothesis, correlational evidence between relevant linguistic and non-linguistic variables can offer support of varying degrees. This is particularly true given the two key confounds that must be controlled for—language relatedness and language contact—and the lack of clear consensus on the various methods available for such controls. In the case of the labiodentals hypothesis, correlational evidence suggests that hunter-gatherers with rougher diets are indeed much less likely to use labiodental phonemes (Blasi et al. 2019). Additional correlational evidence suggests that the languages of hunter-gatherers use labiodental sounds less frequently, even considering only those languages in which labiodental consonants are phonemic. They are particularly uncommon in such languages in word-initial contexts (Everett and Chen 2021). In the case of the aridity hypothesis in Everett (2017), correlational evidence suggests that languages traditionally spoken in regions with low specific humidity rely less on vowels, i.e. have reduced vowel ratios in speech, judging from transcriptions of basic words.

There are various approaches to control for phylogenetic areal influences in such correlational work, and some maintain that the approaches taken to date are insufficient (Hartmann 2022). Given inherent limitations in the typological data, namely that there is a limited number of language families traditionally spoken in very arid regions, it is unclear whether any correlational methods could meet some of the suggested standards. This may lead some to discard correlational approaches entirely, even though they have uncovered associations that support other hypotheses in uncontroversial ways. Consider the correlation, mentioned above, between words for snow/ice and ambient temperature. It could be argued that this correlation too does not control sufficiently for phylogenetic and contact-based factors, given the limited number of language families in cold regions, yet the intuitive nature of the association discussed in Regier et al. (2016) likely makes this correlational support more palatable. Many less obvious associations could exist that reflect direct or indirect causal associations between languages and language-external factors, and in my view, we should be careful not to discard all of these. We can do so while simultaneously acknowledging the inherent limitations of correlational data. Such data can at least demonstrate consistency with the hypothesis generated via step (ii). Of course, many spurious correlations may also exist between linguistic and

9780192886491_Book.indb 132 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:30

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

non-linguistic variables (Roberts and Winters 2013), which is why step (i) is critical to this approach. The method being described here begins not with a correlation, but with a hypothesis that is motivated by well-established results gathered via distinct methods.

Correlational data generated via step (iii) can nevertheless aim to control for phylogeny and contact through varied approaches. One common approach is to use a frequentist or Bayesian linear mixed-effects model in which language family and geographic region are treated as random nested hierarchical effects and the independent variable is treated as a fixed effect. This is the approach used, for instance in Everett and Chen (2021), to demonstrate correlational consistency between agricultural subsistence and increased use of labiodental consonants. Other approaches can also be used, including random sampling from language families and regions (Everett et al. 2015; Everett 2017). None of these approaches is perfect, but each can demonstrate proof of concept necessary to proceed to the next step. This is particularly true when the correlation in question is quite robust to multiple sorts of tests, as evidenced for instance in Everett and Chen (2021).

iv) Assuming correlational data support the hypothesis, other types of data may be gathered. This could include data from historical reconstructions, biomechanical modelling, large-scale analyses of acoustic data, and/or quasi-experimental examinations of speech under distinct external pressures. Note that, to some, diachronic tests may be seen as necessary earlier tests, not something to be done post-correlation-testing. The outline I am offering here reflects my own approach to this topic. Liang et al. (2023) utilize the methodological innovation of a large-scale analysis of acoustic data across 997 locations in China, finding strong support for the hypothesis in Everett et al. (2015, 2016a). Blasi et al. (2019) offer compelling evidence that labiodental phonemes have become more prevalent in Indo-European languages over the last few millennia, alongside the adoption of agriculture and softer diets. In that study, the authors also present biomechanical modelling, supporting the claim that labiodental consonants require less effort for speakers with overbite/overjet dental characteristics. Analogous findings from biomechanical modelling have been offered to support hypotheses on language-external effects on the production of clicks (Moisik and Dediu 2017). In the case of the labiodentals hypothesis, there is also evidence from visual inspections of speech, as discussed in the following section.

7.3 What we have learned: Two illustrations

The approach outlined in the previous section has led to several findings, though the significance of these findings remains disputed. Regardless of one's stance on the

9780192886491_Book.indb 133 /12_first_proofs/first_pr…/xml_for_typesetting 27-Mar-25 23:02:30

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

plausibility of the underlying hypotheses, however, the correlations uncovered in support of them are evident when contrasting linguistic data with external variables. One uncovered association is that between the reduced use of vowels, judging from transcriptions of word lists, and aridity. In Everett (2017), I presented an analysis of the ratio of vowels, as a proportion of all sounds in ASJP word lists, for each of 4,012 dialects (for more on the ASJP database, see Wichmann et al. 2022). Each of these dialects is associated with a location in the ASJP database, and the mean specific humidity of that location was obtained. Regions of low specific humidity are typically cold regions, given that cold air cannot hold much water vapour. As an additional way of illustrating the correlation, we might consider the location of the dialects with the lowest and highest ratios of vowels, by mapping them against the temperature of these locations. In Figure 7.1, I adopt one illustrative approach, plotting languages on a world map with a raster

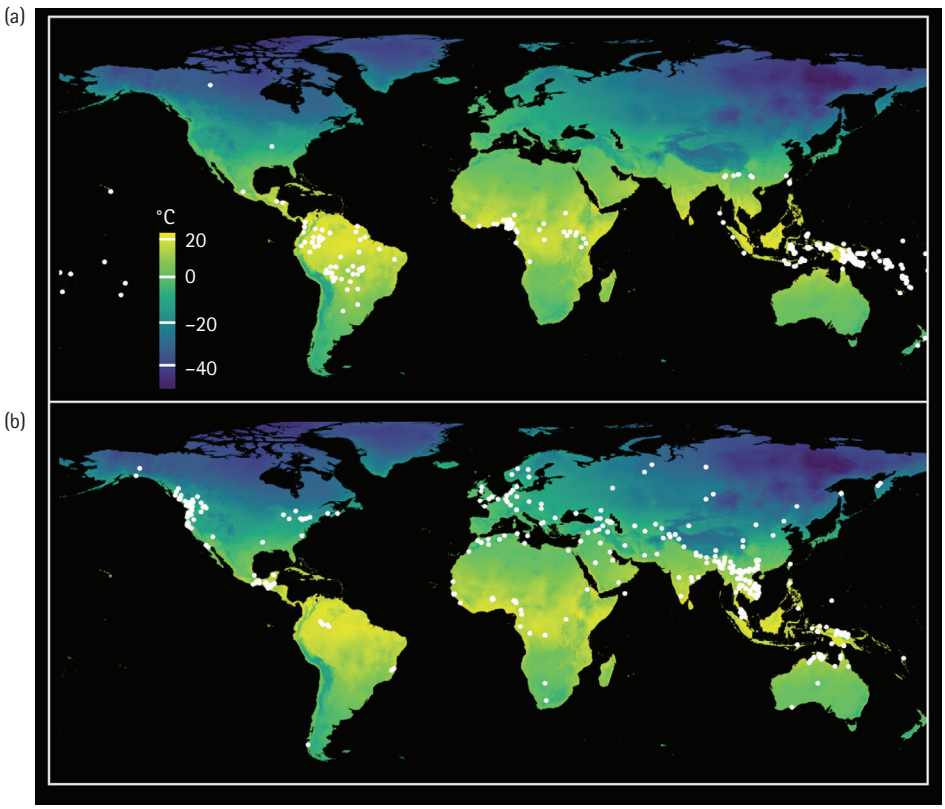


FIGURE 7.1 Two world maps, each with a raster layer based on bioclimatic variable six from the WorldClim dataset (Fick and Hijmans 2017).

This variable represents the mean temperature of the coldest month of the year. Panel A: The location of the top 10 per cent (n = 400) of languages, in terms of reliance on vowels in the ASJP data (Wichmann et al. 2022). Panel B: The location of the bottom 10 per cent of languages, in terms of reliance on vowels in the ASJP data. Dashed lines represent Tropic of Cancer and Tropic of Capricorn. Maps generated via the geom_raster function in R (R Core Team 2021). Vowel data based on Everett (2017).

9780192886491_Book.indb 134 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:31

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

layer based on the mean temperature of the coldest month of the year. In the top panel of the figure, the 10 per cent of languages with the highest vowel ratios (n = 400) are plotted on that layer. In the bottom figure, the 10 per cent of languages with the lowest vowel ratios are plotted. The figure is meant simply to demonstrate the stark contrast of locations between languages with very low and very high vowel ratios. As noted in Everett (2021), these ratios correspond well with other accounts of syllable structures in the relevant language families. If one contrasts the distributions evident in Figure 7.1 with the distribution of languages according to how they encode 'snow' and 'ice', evident in Regier et al. (2016: fig. 1), it is apparent that the distribution of vowel ratios is even more striking.

This clear contrast in Figure 7.1 could nevertheless be due to confounds (see Hartmann 2022). This draws us to an important issue regarding the association between phonation reliance and ambient aridity, and cross-linguistic correlations associated therewith. How does one adequately control for the confounds associated with distributions in such data? As noted above, two well-known confounds are language contact and language phylogeny. One common way to control for these is to use linear mixed methods in which the latter two factors are treated as nested hierarchical factors (see Wang et al. 2023). Under such approaches, language family and geographic region are used as random factors in the model. Two recent studies have adopted more conservative approaches and have suggested that the correlation between ambient humidity on phonation, whether vowel reliance (Hartmann 2022) or tonality (Bromham and Yaxley 2023), is not robust to such approaches. While these are welcome efforts to refine the methods in such work, one could argue that they only interact superficially with the proposed causal mechanism of the hypothesis, given that they only include tests of a general association between humidity and phonation, rather than tests of the effects of very dry air *per se*. Relatedly, it is worth observing that such studies do not include 'sanity checks', wherein the proposed approaches demonstrate that widely accepted environmental effects on human culture/behaviour, of which there are many, are robust to such tests. In the absence of such sanity checks, it is difficult to interpret the significance of the results. It is worth noting that, to some researchers, such approaches may be too conservative, since environmental effects on sound systems may only surface at very large timescales and should not be expected to occur intralinguistically or be completely immune to the effects of relatedness (see discussion in Wang et al. 2023: 3).

Studies in laryngology suggest that only pronounced desiccation yields effects on vocal cord vibration, including increased jitter, shimmer, and increases to phonation threshold pressure (see, for instance, the meta-analysis in Alves et al. 2019). Relevant results are based on *in-vivo* and *ex-vivo* experiments in lab settings, in which ambient water vapour levels were manipulated. Intriguingly, we now have evidence that very dry air impacts naturalistic speech outside the lab, but also that there is no broad association between humidity and speech. This evidence comes in the form of data on jitter and shimmer in the speech of individuals across China. Here, we will focus on jitter. To test for the effects of very dry air on the vocal cords in naturalistic contexts, Liang

9780192886491_Book.indb 135 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:31

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

et al. (2023) examined jitter rates in nearly 1.2 million audio files across 997 locations in China. The recordings were collected with the same methods, sample stimuli, and equipment, with male speakers of roughly the same age.

Jitter is a measure of the imprecision of vocal cord vibration in speech, with higher jitter values corresponding to a more erratic fundamental frequency. It is measured by testing the durations between each vibration of the vocal cords and obtaining an average rate of variation of these durations. Jitter is frequently expressed as a percentage derived from the formula below, through which we can obtain the average difference between consecutive periods of the fundamental frequency in a recording. Liang et al. (2023) used this formula to test the nearly 1.2 million audio files in question, recorded across many distinct regions in China. In the formula, T_i is the duration of the ith interval between vocal cord vibrations and N is the number of intervals tested.

$$jitter = \frac{\frac{1}{N-1}\sum_{i=1}^{n-1}|T_i - T_{i-1}|}{\frac{1}{n}\sum_{i=1}^{n} T_i}$$

Liang et al. (2023) found highly significant associations between ambient aridity and increased jitter, as well as aridity and shimmer (the latter is not discussed here). The strong association between jitter and aridity held despite the fact that neither language contact nor language relatedness should impact jitter rates. In contrast, there is clear evidence in laryngology that very dry air should yield increased rates of jitter. Critically, the association between aridity and jitter obtained in Liang et al. (2023) is driven by very dry regions only, as suggested by the laryngology literature. To illustrate this point, I used the jitter data in the supplemental information of Liang et al. (2023) to create Figure 7.2, which plots jitter averages for the 997 recording locations. In the figure, the locations are grouped into deciles, based on average ambient water vapour. It is clear in the figure that the association in question is driven primarily by the very driest regions. In other words, there is no broad association between humidity and speech, but there is, nevertheless, a clear pattern: very dry air appears to impact phonation in real-world speech, as predicted by the lab-based literature in laryngology.

While the evidence in Figure 7.2 does not demonstrate that the effects of dry air on phonation impact how sound systems evolve, they offer clear evidence of an environmental effect on speech outside the lab. Liang et al. (2023) argue that the data in Figure 7.2, along with other data they provide, strongly support the claims of Everett et al. (2015, 2016a).

This discussion has been focused primarily to the hypothesis related to desiccated ambient air, largely since several recent studies have offered evidence for and against the hypothesis. Moving on to the second hypothesis discussed, related to labiodental usage, there is evidence that hunter-gatherers are less likely to develop overbite and overjet dental configurations as adults. These dental characteristics yield a greater proximity of

9780192886491_Book.indb 136 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:31

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

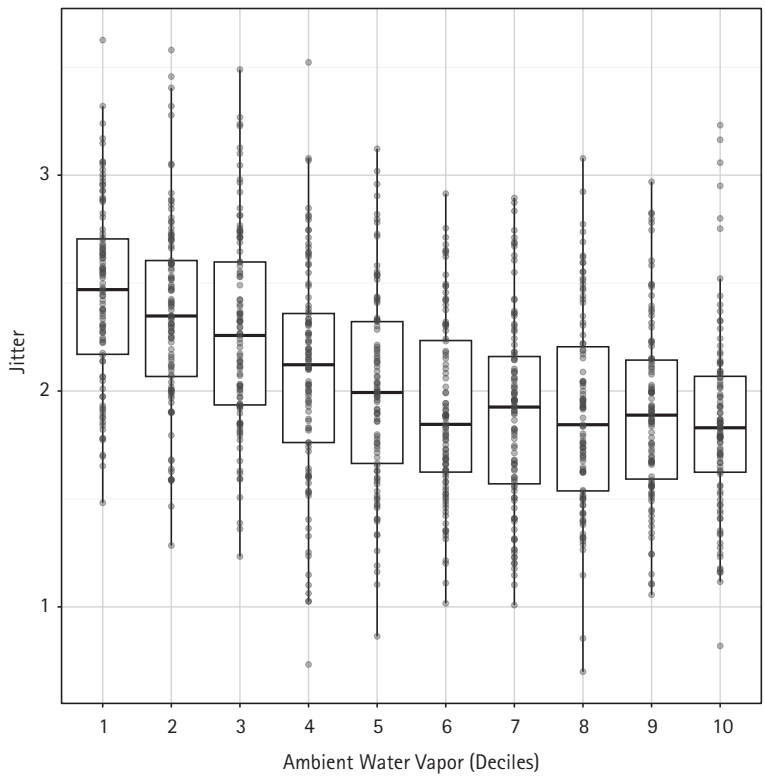


FIGURE 7.2 Average jitter rates (see Formula 1) of recordings in 997 locations in China, in accordance with deciles of average ambient water vapour pressure of those locations.
Based on 1,174,686 recordings.

Source: Data taken from supplemental information in Liang et al. (2023).

the upper lip and bottom teeth in cultures with softer diets, which in turn yields reduced effort of articulation required for labiodental consonants in such populations (Blasi et al. 2019). Figure 7.3 depicts the association between labiodental usage and primary subsistence strategy, across 2,729 populations considered in Everett and Chen (2021). This figure is also derived from transcription data in the ASJP database, and is based on the ratio of labiodental consonants, as a portion of all consonants in each word list (the actual data underlying Figure 7.3 are available in Everett and Chen (2021). Given the pieces of evidence gathered via step (iv) for the labiodentals particular hypothesis, including biomechanical modelling and diachronic reconstructions, the distribution in Figure 7.3 plausibly reflects a typological signal of a way in which the sound systems of languages have subtly adapted to external factors, namely cross-cultural differences in dentition that owe themselves to non-linguistic behaviour, related indirectly to environmental factors.

As in the case of the hypothesis that very dry ambient air impacts the use of the vocal cords across languages, one of the most compelling pieces of evidence related to the labiodentals hypothesis is gathered from actual speech. While actual speech

9780192886491_Book.indb 137 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:33

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

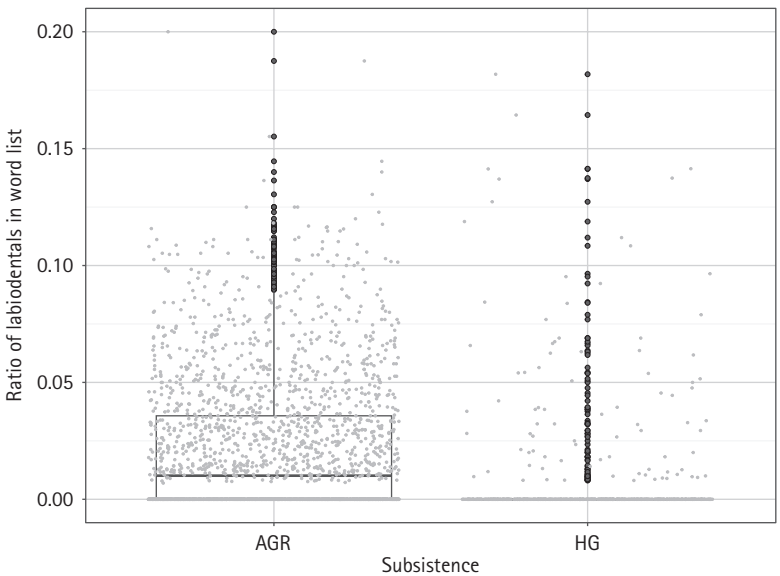


FIGURE 7.3 Labiodental usage across 2,729 dialects, categorized according to dominant subsistence strategy of the associated culture.

Each dot represents a dialect. Within each column, dots are separated along the x-axis via the jitter function in R (R Core Team 2021). The boxplot for HG groups is invisible since the majority have zero labiodentals in their word lists.

cannot demonstrate the putative diachronic trends in question, it can offer evidence for the proximate causal mechanism in question, in this case that variations in bite type impact the ease of use of labiodental consonants. In Everett and Chen (2021), we examined in detail the articulations of ten speakers with a range of bite types, relying on high resolution videos of speaker interviews. Across the ten speakers, it was clear that those with greater overbite/overjet, as judged by others unfamiliar with the hypothesis or results, relied on labiodental consonants more often. In contrast, those speakers with edge-to-edge bites relied on labiodental consonants less than expected, with some labiodental phonemes being produced as bilabials upon close inspection. Conversely, some speakers with pronounced overbite/overjet produced some bilabial phonemes as labiodentals upon careful inspection. In short, evidence from real-world speech now yields further support for the labiodentals hypothesis. Similarly, the work of Liang et al. (2023) offers real-world evidence that ambient aridity impacts phonation in speech.

7.4 Conclusion

It is increasingly clear that languages adapt to external pressures. Least surprisingly and controversially, linguistic adaptation applies to some lexical phenomena. Yet, it also

9780192886491_Book.indb 138 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:34

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

appears to apply to the morphological and phonetic/phonological phenomena, as has become increasingly clear in the last decade. As I have noted in this chapter, the extent of support for phonetic/phonological adaptation varies depending on the hypothesis considered. Nevertheless, the general methodological approach outlined and illustrated in this chapter has offered some support for the notion that the world's sound systems adapt subtly and over the long term to external, non-linguistic pressures.

It is worth stressing that work on this topic is generally quite new, with the bulk of studies having been carried out in the last few years. Some of the hypotheses related to this topic may be discarded with time. Other hypotheses may yet be generated. Finally, some hypotheses may become more widely accepted by language researchers as divergent types of evidence converge in support of them. With time, we should better understand the extent to which the sound systems of languages, like other features of languages and cultures more broadly, adapt to external pressures.

References

Alves, M., Krüger, E., Pillay, B., et al. (2019). 'The effect of hydration on voice quality in adults: A systematic review', *Journal of Voice* 33:1, https://doi.org/10.1016/j.jvoice.2017.10.001.

Bentz, C., Dediu, D., Verkerk, A., and Jager, G. (2018). 'The evolution of language families is shaped by the environment beyond neutral drift', *Nature Human Behaviour* 2(11): 816–821.

Blasi, D., Moran, S., Moisik, S., et al. (2019). 'Human sound systems are shaped by post-Neolithic changes in bite configuration', *Science* 363: 6432.

Bromham, L. and Yaxley, K. (2023). 'Neighbours and relatives: Accounting for spatial distribution when testing causal hypotheses in cultural evolution', *Evolutionary Human Sciences* 5: e27, https://doi.org.10.1017/ehx.2023.23.

Brunelle, J. A., Bhat, M., and Lipton, J. A. (1996). 'Prevalence and distribution of selected occlusal characteristics in the US population, 1988–1991', *Journal of Dental Research* 75: 706–713.

De Boer, B. (2016). 'Is the effect of desiccation large enough?', *Journal of Language Evolution* 1: 55–57.

Dediu, D., Janssen, R., and Moisik, S. (2017). 'Language is not isolated from its wider environment: Vocal tract influences on the evolution of speech and language', *Language & Communication* 54: 9–20.

Dediu, D., Janssen, R., and Moisik, S. (2019). 'Weak biases emerging from vocal tract anatomy shape the repeated transmission of vowels', *Nature Human Behaviour* 3: 1107–1115.

Everett, C. (2013). 'Evidence for direct geographic influences on linguistic sounds: The case of ejectives', *PLoS ONE* 8(6): e65275, doi:10.1371/journal.pone.0065275.

Everett, C. (2017). 'Languages in drier climates use fewer vowels', *Frontiers in Psychology* 8: 1285, doi:10.3389/fpsyg.2017.01285.

Everett, C. (2018). 'The global dispreference for posterior voiced obstruents: A quantitative assessment of word list data', *Language* 94: e311–e323.

Everett, C. (2021). 'The sounds of prehistoric speech', *Philosophical Transactions of the Royal Society B*. doi: 10.1098/rstb.2020.0195

9780192886491_Book.indb 139 /12_first_proofs/first_pr…/xml_for_typesetting 27-Mar-25 23:02:34

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Everett, C., Blasi, D., and Roberts, S. G. (2015). 'Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots', *Proceedings of the National Academy of Sciences of the USA* 112: 1322–1327.

Everett, C., Blasi, D. E., and Roberts, S. G. (2016a). 'Language evolution and climate: The case of desiccation and tone', *Journal of Language Evolution* 1(1): 33–46.

Everett, C., Blasi, D. E., and Roberts, S. G. (2016b). 'Response: Climate and language: has the discourse shifted?', *Journal of Language Evolution* 1(1): 83–87.

Everett, C. and Chen, S. (2021). 'Speech adapts to differences in dentition within and across populations', *Scientific Reports* 11: 1066.

Evteev, A., Cardini, A., Morozova, I., and O'Higgins, P. (2014). 'Extreme climate, rather than population history, explains mid-facial morphology of northern Asians', *American Journal of Biological Anthropology* 153(3): 449–462.

Fan, S., Hansen, M., Lo, Y., and Tishkof, S. (2016). 'Going global by adapting local: A review of recent human adaptation', *Science* 354(6308): 54–59.

Fick, S. E. and Hijmans, R. J. (2017). 'WorldClim 2: New 1km spatial resolution climate surfaces for global land areas', *International Journal of Climatology* 37: 4302–4315.

Gibson, E., Futrell, R., Jara-Ettinger, J., et al. (2017). 'Color naming across languages reflects color use', *Proceedings of the National Academy of Sciences of the USA* 114: 785–790.

Hanel, A. and Carlberg, C. (2020). 'Skin colour and vitamin D: An update', *Experimental Dermatology* 29: 864–875.

Hartmann, F. (2022). 'Methodological problems in quantitative research on environmental effects on phonology', *Journal of Language Evolution* 7: 95–119.

Henrich, J. (2015). *The Secret of Our Success*. Princeton, NJ: Princeton University Press.

Hockett, C. F. (1985). 'Distinguished lecture: F', *American Anthropologist* 87: 263–281.

Ladd, R. (2016). 'Tone languages and laryngeal precision', *Journal of Language Evolution* 1: 70–72.

Leydon, C., Sivasankar, M., Falciglia, D., et al. (2009). 'Vocal fold surface hydration: A review', *Journal of Voice* 23: 658–665.

Liang, Y., Wang, L., Wichmann, S., et al. (2023). 'Languages in China link climate, voice quality, and tone in a causal chain', *Humanities and Social Sciences Communications* 10: 453.

Lupyan, G. and Dale, R. A. C. (2010). 'Language structure is partly determined by social structure', *PLOS ONE* 5: e8559.

Maddieson, I. and Benedict, K. (2023). 'Demonstrating environmental impacts on the sound structure of languages: Challenges and solutions', *Frontiers in Psychology* 14: 1200463, https://doi.org10.3389/fpsyg.2023.1200463.

Meiri, S. and Dayan, T. (2003). 'On the validity of Bergmann's rule', *Journal of Biogeography* 30: 331–351.

Moisik, S. and Dediu, D. (2017). 'Anatomical biasing and clicks: Evidence from biomechanical modeling', *Journal of Language Evolution* 2: 37–51.

Moran, S. and Urban, M. (2021). 'Altitude and the distributional typology of language structure: Ejectives and beyond', *PLOS ONE* 16(2): e0245522, https://doi.org/10.1371/journal.pone.0245522.

Nölle, J., Kirby, S., Culbertson, J., and Smith, K. (2020). 'Does environment shape spatial language? A virtual reality experiment', *The Evolution of Language: Proceedings of the 13th International Conference* 321–323.

Ohala, J. and Ohala, M. (1993). 'The phonetics of nasal phonology: Theorems and data', in M. Huffman, and R. Krakow (eds.), *Nasals, Nasalization and the Velum, Phonetics and Phonology Volume* 5, 225–249. New York: Academic Press.

9780192886491_Book.indb 140 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:34

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Prugnolle, F., Manica, A., and Balloux, F. (2005). 'Geography predicts neutral genetic diversity of human populations', *Current Biology* 15: R159–R160.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/.

Raviv, L., Meyer, A., and Lev-Ari, S. (2019). 'Larger communities create more systematic languages', *Proceedings of the Royal Society B* 286: 2019262.

Regier, T., Carstensen, A., and Kemp, C. (2016). 'Languages support efficient communication about the environment: Words for snow revisited', *PLOS ONE* 11: e0151138.

Richerson, P. and Boyd, R. (2008). *Not by Genes Alone: How Culture Transformed Human Evolution*. Chicago: The University of Chicago Press.

Ringe, Jd. (1998). 'Vitamin D deficiency and osteopathies', *Osteoporosis International* 8: S35–S39.

Roberts, S. (2018). 'Robust, causal and incremental approaches to investigating linguistic adaptation', *Frontiers in Psychology* 9: 2018.00166.

Roberts, S. and Winters, J. (2013). 'Linguistic diversity and traffic accidents: Lessons from statistical studies of cultural trains', *PLOS ONE* 8: e70902.

Shcherbakova, O., Michaelis, S. M., Haynie, H. J., et al. (2023). 'Societies of strangers do not speak less complex languages', *Science Advances* 9(33): eadf7704.

Sivasankar, M. P. and Fisher, K. (2002). 'Oral breathing increases Pth and vocal effort by superficial drying of vocal fold mucosa', *Journal of Voice* 16: 172–181, doi: 10.1016/S0892–1997(02)00087-5.

Sundarrajan, A., Brinton Fujiki, R., Loerch, S., et al. (2017). 'Vocal loading and environmental humidity effects in older adults', *Journal of Voice* 31(6): 707–713, doi: 10.1016/j.jvoice.2017.02.002.

Vilela, B., Fristoe, T., Tuff, T., et al. (2020). 'Cultural transmission and ecological opportunity jointly shaped global patterns of reliance on agriculture', *Evolutionary Human Sciences* 2: E53.

Wang, T., Wichmann, S., Xia, Q., and Ran, Q. (2023). 'Temperature shapes language sonority: Revalidation from a large dataset', *PNAS Nexus* 12: 384, https://doi.org/10.1093/pnasnexus/pgad384.

Wichmann, S., Holman, E., and Brown, C. H. (eds.). (2022). The ASJP Database (version 20).

9780192886491_Book.indb 141 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:34

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

9780192886491_Book.indb 142 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:34

CHAPTER 8

THE EVOLUTION OF THE GRAMMAR OF THE BODY IN SIGN LANGUAGES

WENDY SANDLER, SVETLANA DACHKOVSKY, AND ROSE STAMP

8.1 Introduction

The days of studying language and its origins through written transcriptions are over. Sign languages teach us how valuable the visual signal is, and today, with the widespread and accessible use of video, and a new appreciation of visible gestures, only video recordings of any language can serve as data for the investigation of language evolution.

We will never know whether a visible gestural system preceded speech in language evolution (see Corballis 2003; Arbib 2012), but there is no doubt that both vocalization and visible gestures persist in human communication to this day. Chimpanzees, descended from a common ancestor of humans millions of years ago, and thus our closest living evolutionary relatives, also make use of both vocalization and gesture in their communication (Liebal et al. 2014; Oña et al. 2019). These hard facts make it reasonable to assume that both vocalization and gesture were at least present from the get-go.

The existence of contemporary sign languages provides a unique methodological tool for identifying the core of human language, for two reasons. First, in sign languages, actions of visible articulators of parts of the body correspond to specific linguistic functions. It is the body that reflects the way information is structured at all levels of language (Sandler 2018). We propose that this visible display in sign languages is an accessible clue to the structuring of spoken language as well, structuring that is not directly observable in speech. In other words, linguistic units and structures that must be inferred in spoken languages are explicitly accessible in sign languages. Second, in

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

contemporary times, only sign languages emerge and structure themselves *de novo*, and this emergence can be observed and traced in real time. Of course, contemporary deaf people share fully evolved human brains with the rest of us, and we cannot claim that the emergence of new sign languages mirrors the biological evolution of language. But the newborn natural languages of deaf people provide unique and accessible empirical insight into the way a language evolves in a group of humans. Metaphorically speaking, sign languages offer fast forward evidence of language evolution, if one knows where to look.

Our work on emerging sign languages has yielded a methodological tool: documenting the recruitment of the body for linguistic expression. By attending to the emergence of systematic articulations of the body, we gain unique insight into the way linguistic structuring emerges in a group of humans.

The definition and identification of linguistic functions in sign languages rely on hundreds of years of research on spoken language. Although we do not have the space to elaborate this point, it is a critical part of our approach, since it provides a common frame of reference for the definition and behaviour of specific linguistic elements and levels of structure (Sandler and Lillo-Martin 2006; Pfau et al. 2012).

Our lab has studied the emergence of two sign languages that developed under different social conditions in present-day Israel, each less than a hundred years old. One is Al-Sayyid Bedouin Sign Language (ABSL), a Village Sign Language, which arose in a relatively insular Bedouin village in the Negev desert. The other is Israeli Sign Language (ISL), the widespread Deaf Community Sign Language of Israel, which emerged at about the same time as ABSL, but developed under conditions that can be compared to those of pidgin-creole formation (McWhorter 2022), and that spread through use in schools, interpreting, and other social frameworks. Here, we highlight the role of the body in rigorous analysis of language emergence, with examples from our own research.

We describe how prosody structures information in sign language in section 8.2. In section 8.3, we turn to ABSL and deal with the property of word order for identifying referents in a sentence in 8.3.1. In our study of this phenomenon across age groups[1] in ABSL, we found that the most common sentence word order in ABSL is Subject-Object-Verb, SOV (Sandler et al. 2005).[2] In section 8.3.2, we provide evidence for mapping the chronological recruitment of bodily articulators to the emergence of linguistic complexity. We illustrate this by showing how two clauses in a complex (conditional) sentence are connected in younger ABSL signers. In the exposition, we note that iconicity afforded by the body contributes to the genesis of words. But this section also demonstrates that understanding words is not enough: only careful observance of bodily actions ensures correct analysis of the data.

[1] The first four signers of ABSL were siblings in the same family. Because of the demographics of the community, in which siblings in the same family can be twenty years apart in age, we found it expedient to refer to each age group documented as a stratum rather than a generation (Sandler 2012).

[2] Later work (Meir et al. 2017) on broader data including silent gesture supported the idea that SOV was most prominent in sentences with two human arguments.

9780192886491_Book.indb 144 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:34

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

We then turn, in section 8.4, to ISL, a young Deaf Community Sign Language that arose under very different social conditions from those of ABSL (Meir et al. 2013; Meir and Sandler 2019). ISL came together as a unified language in a school for deaf children that opened in Jerusalem in 1932. Although the school policy was strictly oral, the children conversed among themselves in sign language. The principal and some of the teachers had come from a deaf school for Jewish children in Berlin, as did some deaf adults who had emigrated to Israel in those years, and who assumed leadership roles in the fledgling deaf community (Meir and Sandler 2007). As a result, certain lexical items made their way into ISL from German Sign Language. This influence notwithstanding, the linguistic structure of ISL developed on its own course, documented by members of our team.

We begin our description of the emergence of ISL in section 8.4.1 with a demonstration of bodily signals in a sentence with a relative clause in the contemporary language. We then describe a rigorous study in section 8.4.2 that traces and quantifies how relative clauses emerged in that language, across generations (Dachkovsky 2018, 2022). In section 8.4.3, we show that careful tracking of the use of the body reveals the ways in which parts of the body become organized over time to reflect the hierarchical structure of a discourse. These studies are united by the Grammar of the Body as a methodological tool for investigating sign language emergence. We propose that the visual salience of sign language articulators, and their direct correspondence to demonstrably linguistic functions, can shed light on the course of emergence of natural languages in general.

Culling the investigations of many sign language researchers in different sign languages over the past several decades (see articles in Pfau et al. 2012), a Grammar of the Body in sign languages is schematized in Figure 8.1—here, from contemporary ISL. The figure shows the correspondence between bodily articulators and the specific linguistic functions associated with the activation of each one.

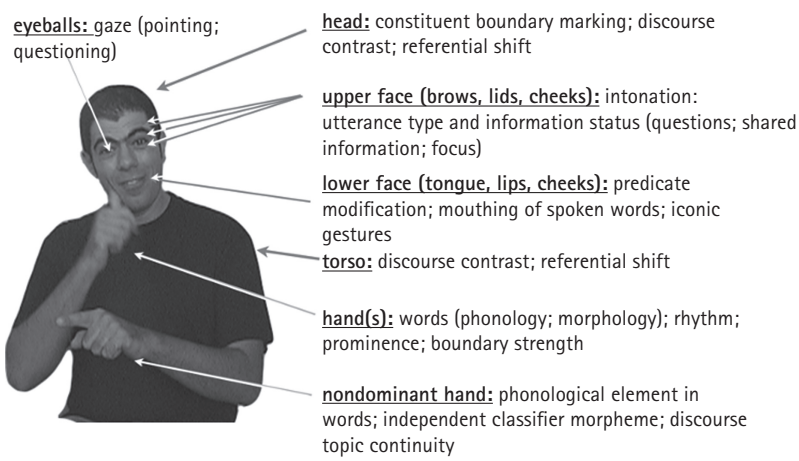


FIGURE 8.1 The Grammar of the Body: some of the linguistic properties conveyed by actions of particular articulators of the body.

Source: From Sandler (2012, 2018), reproduced with permission.

9780192886491_Book.indb 145 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:35

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

8.2 Prosody structures information

When communication is exclusively visual and is conveyed by a large number of articulators whose movements are directly perceivable and often simultaneous, the result is a system that can be both complex and transparent at the same time. This transparency often reveals general linguistic properties that are opaque or covert in spoken languages. In sign languages, complex linguistic composition can be seen at a glance.

It is an established fact that prosodic constituency typically corresponds to syntactic constituency. That is, prosodic breaks typically demarcate syntactic constituents (Selkirk 1984; Nespor and Vogel 1986). However, syntax and prosody are not isomorphic—prosodic breaks sometimes follow information structure in a particular context, and, for this reason, they do not always correspond to syntactic constituents in spoken languages (Nespor and Vogel 1986) or in sign languages (Nespor and Sandler 1999; Sandler 2010). Young languages, such as spoken pidgins and emerging sign languages, may not yet have developed well-defined grammatical structures (like embedded clauses) or function words (like *that*, *which*, *if-then*, *or*, etc.). However, prosody—rhythm, intonation, and stress—seems to be present at the outset. It is often intonation that distinguishes one type of rhythmically demarcated prosodic constituent from another, e.g. conditionals from relative clauses. In sign languages, prosodic signalling of all types is clearly articulated by movements of different parts of the body. Such prosodic signals arise early to mark complex structures, but we find that even they take time to organize and to distinguish different types of constituents. This allows us to track the emergence of linguistic complexity by closely attending to the Grammar of the Body in sign languages (Sandler 2018).

Regarding prosody, very important to our analysis, we take a contemporary approach, which includes visual signals. This is in contrast to historical methods, which rely exclusively on written data, fostering an approach in which prosody is secondary, and is somehow superimposed on linguistic structure. Videotaped data foster a different perspective, one in which the prosody itself organizes information at all levels of structure (Dachkovsky et al. 2022). In sign languages, prosody is expressed by particular and systematic articulations of different parts of the body. Once this salient prosodic system is understood, it is possible to investigate associated levels of linguistic structure, be they lexical, syntactic, or discursive.

Criteria for determining prosodic constituency were developed in a study on ISL (Nespor and Sandler 1999). In that study, major prosodic breaks (intonational phrases) were systematically marked by a combination of manual and non-manual phonetic cues. Three distinct manual cues were found to mark prosodic breaks: (a) holding the hands in place; (b) pause and relaxation of the hands; or (c) repeating the final sign in the constituent. Non-manual cues at the intonational phrase boundary included both (a) a clear change in head or body position, and (b) a concomitant change in facial expression, the latter interpreted by the researchers as sign language intonation

9780192886491_Book.indb 146 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:35

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

(see also Sandler 1999; Sandler and Lillo-Martin 2006). Eyeblink also correlates reliably with intonational phrase boundaries in ASL (Baker-Shenk and Padden 1978; Wilbur 1994) and in ISL (Nespor and Sandler 1999). These breaks separate major prosodic/syntactic constituents, such as topics, extraposed elements, and non-restrictive relative clauses, from the rest of the sentence, and they separate sentences from one another. Just as phrase final lengthening and intonation excursions characterize Intonational Phrase boundaries across spoken languages generally, we proceeded on the assumption that the main rhythmic and intonational cues associated with prosody in a sign language would correlate with such signals.

Prosodic constituency often reflects syntactic constituency; for example, syntactic phrases typically occupy phonological phrases (also called intermediate phrases) and syntactic clauses typically occupy intonational phrases. Although syntax and prosody are not isomorphic, and each is understood to be a separate level of linguistic structure (e.g. Nespor and Vogel 1986), prosody proves useful in determining syntactic structure in the analysis of young languages. Prosody offers a tool for merging strings of words into a single constituent, as well as for breaking them up into separate constituents, and for indicating relations among the constituents. In the next section, we show how this works for the Village Sign Language of Al-Sayyid. Some of the functions of bodily cues that are not prosodic are also important methodologically, as we explain in what follows.

8.3 Al-Sayyid Bedouin Sign Language

Al-Sayyid Bedouin Sign Language (ABSL) began with the birth of four deaf siblings into a single family in the 1930s. Due to genetic reasons (Scott et al. 1996) and consanguineous marriage practices, deafness spread in the village, so that in 2004 there were already about 130 deaf people in a population of 4,000 (Kisch 2004), perhaps the highest proportion of deaf people in any population in the world.

8.3.1 Prosodic structure and word order in ABSL

Constituency within sentences was determined by careful observation of prosodic cues by a research team based in the Sign Language Research Lab at the University of Haifa. The main researchers on this project were Sandler, Meir, Aronoff, and Padden (henceforth, SMAP). Major breaks in the utterance were identified by shifts in the rhythm marked by a pause or lowering of the hands, together with a change in head or body position and facial expressions. These same prosodic cues to major constituent breaks occur in ISL (Nespor and Sandler 1999), and have been observed in other sign languages as well. We now turn to examples of the way in which these prosodic criteria were applied in the analysis of ABSL.

9780192886491_Book.indb 147 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:35

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

One method that was adopted was asking signers to sign video clips in their language to another ABSL signer of about the same age. The SMAP researchers developed 'The Haifa Clips' battery for the elicitations, a video battery that other researchers have also found useful (Ergin et al. 2018). Clips usually involved subjects, verbs, and objects, in events such as a woman giving a shirt to a man, a man washing dishes, etc., and the order was typically Subject-Object-Verb (SOV), e.g. WOMAN SHIRT GIVE; MAN DISH WASH (Sandler et al. 2005; Padden et al. 2010). In addition to access to the stimuli, the meaning of signs is often transparent due to an iconic relation between form and meaning afforded by the body, shown for an older second-generation (Stratum II) signer in Figure 8.2. But even iconicity is influenced by culture, as seen in the sign for MAN, indicating the moustache traditionally sported by Bedouin men.[3]



FIGURE 8.2 Iconic signs in a typical SOV sentence, MAN DISH WASH.

Careful attention to prosodic cues is essential for accurate analysis of young languages. One example of a potential error will suffice to bring this point home. In one narrative, the signer described saving money to build a house and then he signed BUILD WALLS DOORS. This sequence might have been coded as SVO(O)—a counterexample to the SOV order we found generally. But prosodic analysis showed that the signer executed a major prosodic break between BUILD and WALLS, DOORS, indicating that one utterance ended with BUILD, and that it was followed by a sentence fragment, WALLS, DOORS. The translator's voiceover confirmed the prosodic analysis: 'And I began to build. Walls, doors…'

8.3.2 The emergence path of the Grammar of the Body in ABSL

In addition to elicited data, the researchers used spontaneous narratives, and found that the next age group recruited the body to provide more linguistic complexity.

[3] While iconicity and culture are important contributors to early signs, they can lead to different symbols for the same concept, and indeed conventionalization of signs has a life of its own, even in a small community (Sandler et al. 2011).

9780192886491_Book.indb 148 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:35

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

An older second-generation signer (Stratum II) produced a sentence with conditional meaning, a meaning that was understood only from context but that was not clearly marked prosodically. The sentence refers to a child that had received a now defunct folk immunization against scorpion bites. It means, 'If a scorpion stings the child, there will be no effect'. The end of the 'if' part and the beginning of the 'then' part of the sentence are shown in Figure 8.3.



FIGURE 8.3 The boundaries of two clauses with conditional meaning: SCORPION-STING/NO-EFFECT—'If a scorpion stings the child, there will be no effect'.

By contrast, a younger second-generation (Stratum III) signer and other signers of this stratum added head movement and facial expression to convey complex sentences (see also Sandler et al. 2011). In Figure 8.4, we see an illustration of a Stratum III signer signing a conditional sentence. The context is that the signer will talk to the school principal to find out if she can be a teacher's aide there. The string itself is HE REFUSE NOTHING. The 'if' clause is marked prosodically for dependency by head down and brow raise, and the 'then' clause by a complete change in head, face, and body postures. It is the prosody (together with the context) that gives the string the intended conditional reading in a complex sentence, 'If he refuses, then there's nothing I can do'.



FIGURE 8.4 Prosody marks a conditional sentence in a younger second-generation signer of ABSL. REFUSE/NOTHING. 'If he refuses, then there's nothing I can do'.

Source: Sandler et al. (2011).

9780192886491_Book.indb 149 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:36

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

In a preliminary study of representatives of different generations of ABSL signers, each stratum (age groups about twenty years apart) recruited an additional part of the body for linguistic functions in their narratives, and, in so doing, added complexity to the language (Sandler 2012). This emergence path is shown in Figure 8.5.

Stratum	HANDS	HEAD	FACE	BODY	NONDOMINANT HAND
I	X				
II	X	X			
III	X	X	X		
IV	X	X	X	X	X

Stratum	WORDS	COMPLEX SENTENCES	DISCOURSE REFERENCE/COHESION
I	Signs		
II	Signs	– Unsystematic clause linking	
III	Signs	– Complex sentences – Embedding	– Illocutionary force – Parentheticals – Referential shift
IV	Signs	– Complex sentences – Two degrees of embedding	– Illocutionary force – Parentheticals – Double Referential shift – Contrasting two referents in a discourse – Backgrounding

FIGURE 8.5 Recruiting additional articulators adds complexity as a sign language emerges. Data from ABSL.

Source: Sandler (2012, 2013); reproduced with permission.

8.4 Israeli Sign Language: Another young sign language

We now turn to Israeli Sign Language (ISL). It is about the same age as ABSL, but has emerged under very different social conditions. ISL is what we call a Deaf Community Sign Language, used by most deaf people in Israel—about 10,000 people—on a daily basis, as well as in schools and in sign language interpreting. It began life as a kind of creole in the 1930s (Meir and Sandler 2007), and our lab team has been able to track the emergence of linguistic structure by documenting and quantifying the signing across generations of the community.

As was the case with the ABSL studies, our method exploits recruitment of body articulators together with Labov's (1963) Apparent Time Construct, which predicts that

9780192886491_Book.indb 150 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:36

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

the language of older people reflects the state of a grammar when that person was young. Their language serves as a source of comparison with younger language users to determine language change, a prediction Labov later confirmed empirically (Labov 1966).[4]

8.4.1 A standard relative clause in contemporary Israeli Sign Language

If we consider the actions of the body in sign language, and work from body to linguistic structure, general properties of language stand out in high relief. The articulators each mark different linguistic functions, and they are physically independent of one another, which is also an advantage for analysis. This independence makes it possible to incorporate a good deal of simultaneity of structure in sign language utterances, where spoken languages are much more confined to linearity.

Figure 8.6 shows a sentence that means, 'The little dog that I found last week ran away'. Here is the gloss, in which 'IX' stands for 'index', typically a pointing pronominal sign; subscript 'I' stands for an intonational phrase; and subscript 'Ø' stands for a more minor, phonological (or intermediate) phrase: [[DOG SMALL IX] $_{Ø}$ [WEEK-AGO I FIND IX] $_{Ø}$] $_{I}$ [[ESCAPE] $_{Ø}$] $_{I}$

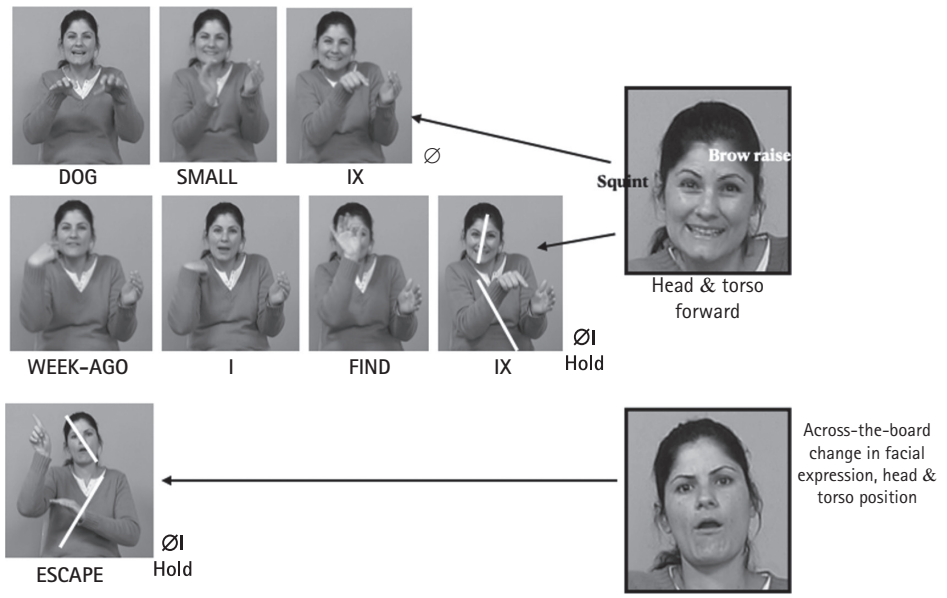


FIGURE 8.6 A complex sentence with a relative clause in ISL, '[The little dog [that I found last week]] ran away].'

Source: Sandler (2018); reproduced with permission.

[4] Labov's Apparent Time Construct is consistent with the Critical Period Hypothesis put forward and supported mainly by brain research in Lenneberg (1967).

9780192886491_Book.indb 151 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:37

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

The sentence illustrates the Grammar of the Body at the sentence level clearly, including prosodic as well as non-prosodic markers.[5] The sentence contains two intonational phrases, separating the topic of the sentence from the comment. The first intonational phrase consists of two lower-level phonological phrases (see Nespor and Sandler 1999). The phrases are signalled by the timing of the hands, and the facial intonation aligns itself with these prosodic constituents, as is the case with intonation patterns in spoken language. The non-dominant hand, configured for SMALL (DOG) serves as a (non-prosodic, lexical) topic continuity marker, persisting in the signal as long as the small dog is the topic, and changing for the comment, ESCAPE ('ran away').

8.4.2 The emergence of the relative clause in ISL

Figure 8.6 above displays a subordinate construction: a relative clause. The emergence of subordination represents one of the central questions in contemporary linguistics. Many linguists have hypothesized that subordinate constructions start off diachronically with prosodic signals, which might precede morphosyntactic markers of subordination (e.g. Mithun 2009; Givón 2015). Contemporary spoken languages cannot provide a testing ground for this hypothesis because all of them have fully grammaticalized subordinate constructions. However, a young sign language, such as Israeli Sign Language (ISL), is perfectly suited for this purpose (Meir and Sandler 2007). In ISL, relative clauses are usually marked by squinted eyes and a forward head position—markers which have been shown to perform the functions of intonation in sign languages (Sandler 1999; Dachkovsky et al. 2013).

This systematic prosodic marking of relative clauses is characteristic of the younger signers' language, but not of the language of the older signers. This suggests that relative clause marking developed over time in the language, which led us to ask just how it emerged. To this end, the ISL data were collected from three age groups of signers through an interactive picture task designed to elicit relative clauses. In order to deduce and evaluate changes in the degree of clause subordination, intonational signals (facial and head movements) were analysed along three parameters of change: (1) degree of obligatoriness; (2) functional change; and (3) phonetic strength. The study adopted Labov's (1963) Apparent Time Construct (see section 8.3 above) to infer diachronic changes from synchronic data collected from the different age groups.

Three people participated in the elicitation task: a deaf signer, a deaf partner, and a native ISL signer who conducted the experiment. In the elicitation task, two identical characters, distinguishable only by their participation in different events (e.g. a man watching TV vs. a man driving a car), and then engaging in an additional activity (e.g.

[5] Although signs can appear to be holistically iconic, Stokoe (1960) first demonstrated for ASL that sublexical units also function as meaningless and contrastive in sign languages. This is true of ISL (Sandler 2010, 2017), and other sign languages as well, manifesting the property of duality of patterning (de Boer et al. 2012).

9780192886491_Book.indb 152 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:37

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

the man combing his hair in addition to watching TV, vs. the man smoking in addition to driving a car) were gradually introduced to the deaf signer. Afterwards, a native deaf experimenter asked the signer a double question about the observed situation (e.g. *Which man is combing his hair and which man is smoking*?). The signer's answer to this double question (viz., 'The man who is watching TV is combing his hair'; 'The man who is driving a car is smoking') enabled the deaf partner to choose the appropriate picture from a set of four cards containing mixed combinations of the referents and their actions. Due to the double question, the expected signed answer was the equivalent of a restrictive relative clause embedded into a full main clause 'The man who is watching TV is combing his hair, and the man who is driving a car is smoking'.

The results of the detailed analysis of face and head movements demonstrated that these signals begin as unsystematic, pragmatic devices, or, more specifically, as information structuring devices. To illustrate this, let's take the string that means 'The man who is driving a car is smoking'. In the older group, forward head movement aligns with the simple nominal phrase 'the man', marking it as an information anchor, or topic, of the whole utterance (Sankoff and Brown 1976; Fox and Thompson 1990) (Figure 8.7). As the communicative situation implies competition between the referents, even at this early stage in the language, the information status of the utterance topic 'man' is marked by squint as not automatically accessible (Ariel 1991; Dachkovsky and Sandler 2009). The hands produce a rhythmic break between the topic, 'the man', and the specifying predication 'driving a car'. However, subordination requires asymmetry between the main clause and the subordinate/modifying clause. Here, the modifying clause 'driving a car' does not co-occur with any intonational cues which signal clausal asymmetry, i.e. linking it to 'the man' or distinguishing it from the main predication 'is smoking'. One could interpret this utterance as a sentence fragment followed by coordinated predicates: 'A man. (He is) driving (and he is) smoking'. In sum, at the older stage, no asymmetry is signalled between functionally subordinate and main clauses.



FIGURE 8.7 Three prosodic chunks in the ISL older signer's response corresponding to the nominal 'the man', the specifying predication 'who is driving a car', and the main predication 'is smoking' in the sentence 'The man who is driving a car is smoking'. The intonational signals are co-temporal only with the topic of the utterance, 'man'.

9780192886491_Book.indb 153 /12_first_proofs/first_pr[illegible]/xml_for_typesetting 27-Mar-25 23:02:37

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

In the middle age group, and even more so in the youngest age group's production, the intonational signals in relative clauses became very systematic both in terms of their frequency and their distribution. More specifically, in the two younger groups, the scope of the forward head movement and squint spreads to include both the nominal 'the man' and the modifying relative clause 'who is driving a car', which merge into one prosodic unit, characterized by forward head movement and squint (Figure 8.8). The merge is underscored by signing DRIVE at the same location as MAN, near the right side of the signer's body. The prosodic boundary between the new nominal constituent with the relative clause, 'man who is driving' and the main predication 'is smoking' is characterized by a complete release/change of the intonational signals. This salient contrast in the bodily signals neatly reflects the clausal asymmetry, which lies at the heart of subordination (Croft 1995; Cristofaro 2005).



FIGURE 8.8 The same string in a younger ISL signer. The nominal and the relative clause, 'The man who is driving a car' constitute one prosodic constituent, marked by the intonational signals, forward head movement, and squint. The main predication, 'is smoking', is separated prosodically from the preceding material by a rhythmic break (marked by a dotted line), as well as by an across-the-board change in intonational signals of the face and head.

In short, the later stages are characterized by a clustering of several intonational and prosodic cues, which mark internal cohesion of the newly emerged relative clause construction, as charted in Figure 8.9. In the process of gradual emergence of relative clauses, intonational face and head cues spread their scope from the simple noun phrase to a complex nominal constituent including a relative clause. As a result, they grammaticalized from more general signals of information structure—such as topic, and not automatically shared information—into signals of clause-level dependencies. Close analysis of face and head movements provides crucial cues to the emergence and development of linguistic complexity in general, and to subordination in particular.

9780192886491_Book.indb 154 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:37

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

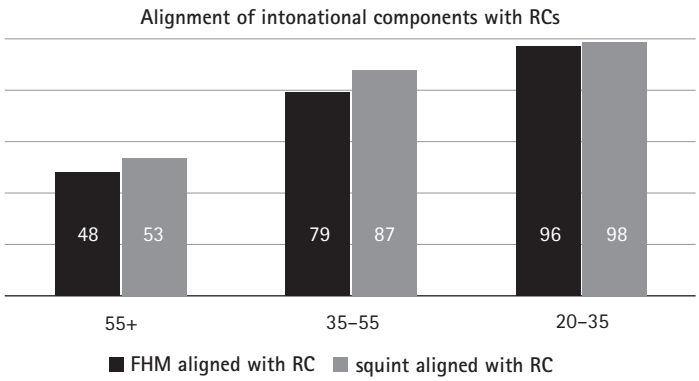


FIGURE 8.9 Changes in the alignment of forward head movement (FHM) and squint with relative clauses across the three age groups of ISL signers.

Note: Intonational signals can occur with or without a prosodic reset after the nominal head (e.g. *the man*) both in spoken (Birkner 2012) and in signed languages (Dachkovsky 2022). This fact has important implications for the analysis of the function of the intonational components, but is beyond the scope of the present chapter.

8.4.3 Mapping the body onto the discourse hierarchy in ISL

It is not only at the sentence level that bodily cues reorganize diachronically to mark linguistic structure. By studying the systematic actions of the body at the level of discourse as well, we were able to track the way the articulators of the body reorganize across generations of ISL signers to reflect different levels of the discourse hierarchy. This method, which adopted a particular theory of discourse structure, was extremely illuminating, allowing us to distinguish a signer's cognitive structuring of a discourse from the means for expressing it. We found that cognitive structuring of a discourse is present but not marked in the oldest group, while systematic coordination of bodily actions for expressing this structure emerged over time in the language, and is present in the younger signers (Dachkovsky et al. 2022).

Specifically, we investigated whether the conceptualization of hierarchical organization is present before its overt linguistic structuring. For this, we turned again to the body as our point of analysis, looking at the different ways in which the body, through movements of the face, head, and torso, can signal boundaries and relations at a discourse level. Here, we concentrated both on prosodic signals, such as those described in the previous sections, as well as on the size and combinations of articulators at different levels of the discourse hierarchy. We adopted the Rhetorical Structure Theory, RST (Mann and Thompson 1988), as a tool for parsing our signed texts.

When we communicate, we organize our text in a hierarchical way (Hobbs 1979; Grosz and Sidner 1986; Mann and Thompson 1988; Polanyi 1988; Van Kuppevelt 1995), marking the relation between different units of the text. The use of RST allowed us to

9780192886491_Book.indb 155 /12_first_proofs/first_pr…/xml_for_typesetting 27-Mar-25 23:02:38

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

segment each signed narrative into its smaller units based on the semantic relations which hold between them. An entire signed text can be represented visually as a tree showing all the relations between units as well as the hierarchical nature of the narrative (as shown in Figure 8.10 below). The tree shows how the text is parsed into smaller units. For example, at the highest level of the hierarchy, the entire text can be broken down into two segments connected by a single RST function, Summary, which connects the summary in lines 50–56 to the rest of the text in lines 1–49. Smaller units, like segments of Elaboration or Evidence, appear at lower levels of the hierarchy.

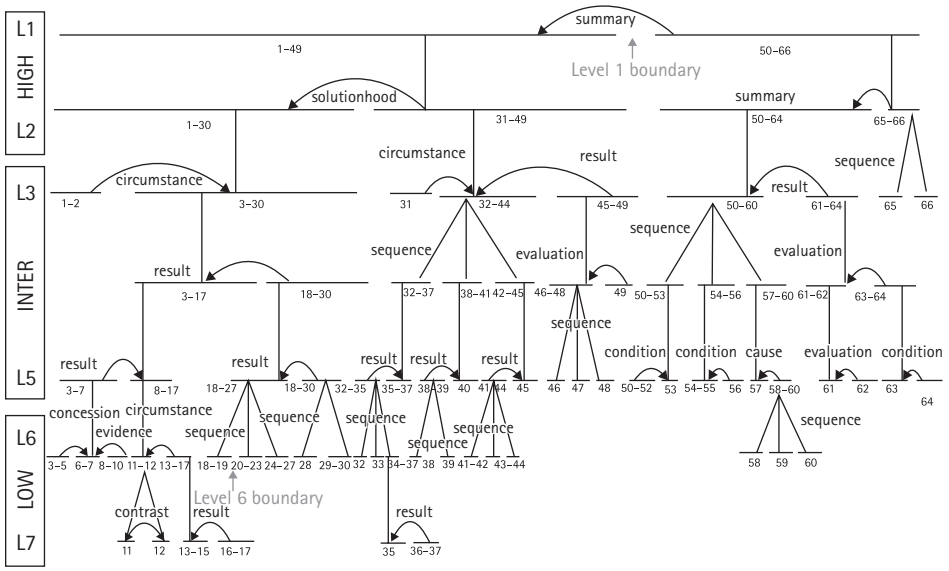


FIGURE 8.10 Example of RST structure produced by an older ISL signer, analysed conceptually, with higher levels shown at the top and lower levels shown at the bottom.

The use of RST enabled us to characterize the conceptual depth of a narrative—the depth of the conceptual information expressed—and to examine the overt linguistic marking at various levels of the text structure. The conceptual depth of the narrative is represented by the number of levels of depth expressed through the conceptual organization semantically and pragmatically, regardless of whether it is overtly marked or not. After parsing the texts of ten deaf ISL signers, we divided them into two age groups: five signers over the age of 55 (older) and five signers between the ages of 18 and 54 (younger). We minutely coded all overt linguistic markers at each boundary of the narrative hierarchy that we had determined based on meaning, according to the RST methodology.

Crucially, we found the same level of conceptual depth in both age groups. All participants, young and old, produced narratives with equally deep conceptual structures. Texts from both older and younger signers reached a depth of seven levels, as shown in Figure 8.10 above, which represents an example of a narrative produced

9780192886491_Book.indb 156 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:38

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

by an older ISL signer. However, all narratives demonstrated a gradual and systematic decrease in the number of cues based on the level of the RST hierarchy, suggesting that differences between discourse levels are predictable and systematic from the very beginning of language emergence. This finding also follows studies of spoken language (e.g. Mayer et al. 2006), which show that boundaries at higher levels of the discourse tend to be marked by stronger prosodic cues than boundaries at lower levels. Fenlon et al. (2007) showed very similar findings in British Sign Language.

Our study of a young sign language showed that different levels of the discourse hierarchy are already present **conceptually**, but that signalling them emerges gradually. In other words, although there is no difference in conceptual depth across age groups, there are clear differences in the expression of this structure (Dachkovsky et al. 2022).

For example, older signers used large articulators such as the torso at both high and low levels of the discourse hierarchy. At the highest level, an older signer, shown in Figure 8.11, produces a strong torso movement forward, dividing the narrative into its two largest segments, where the boundary lies between the two units of Summary (shown on Figure 8.10). The presence of torso movement, however, is also seen at the lower-level boundary of a relation that connects a sequence of three related expressions, described as 'Sequence' according to RST (see Level 6 boundary on Figure 8.10). In this example of an older signer, torso and head movements accompany each individual boundary, enacting the situation mimetically. Older signers use their bodies differently, often exploiting the use of the torso mimetically rather than in accord with the level of the discourse hierarchy.

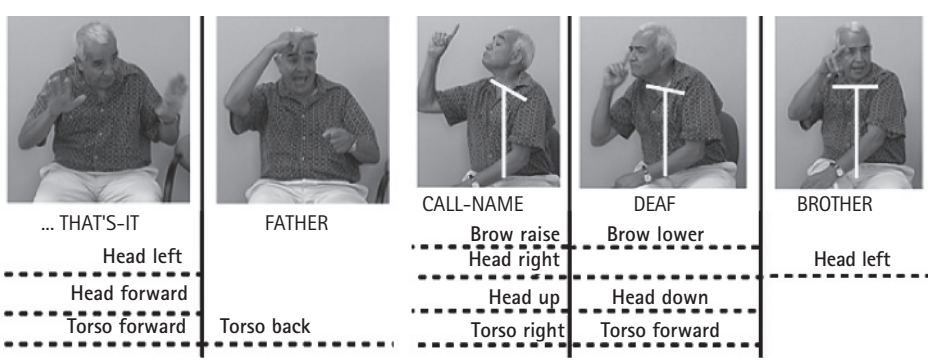


FIGURE 8.11 Older signer: a higher-level boundary (Level 1, Summary) is signalled through the use of a lexical marker THAT'S-IT as well as head and torso movements. At the lower-level boundary (Level 6, Sequence), the signer also marks the breaks between three related sequenced expressions with head and torso movements.

Note: The figure presents the signs at the boundary of a relation but omits the rest of the lexical material due to space limitations.

Source: from Dachkovsky et al. (2022); reproduced with permission.

In contrast, only younger signers mark higher levels of the discourse structure with larger articulators, showing a systematic mapping between the level of the discourse

9780192886491_Book.indb 157 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:38

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

unit and the size of the articulator. As demonstrated in Figure 8.12 below, for a younger signer, while a higher-level boundary is signalled with multiple cues, including a clear torso movement forward, a lower-level boundary is signalled by a change only in head and facial movements, without any torso movement. As a result, it is only for younger signers that torso movements are the most reliable indicator of the difference between discourse levels.

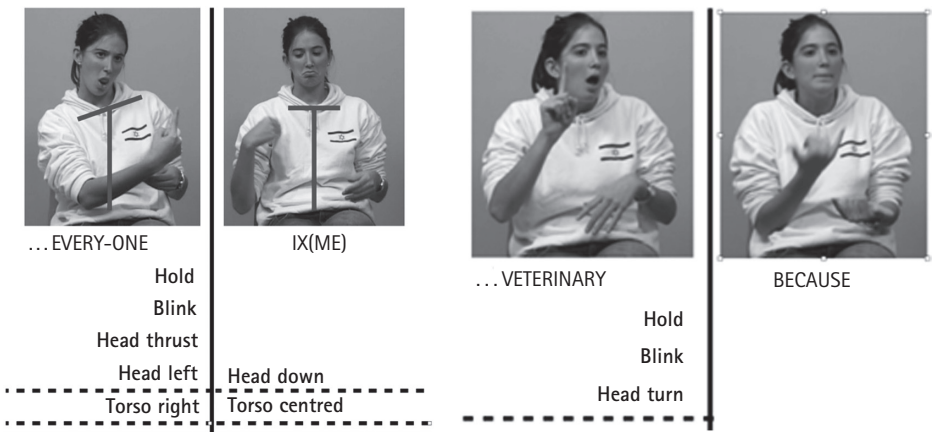


FIGURE 8.12 Younger signer: The higher-level boundary (Level 2, Circumstance) (after EVERYONE) is signalled by multiple cues including a change of torso position, head thrust, and blink. At the lower level (Level 5, Cause), the signer explains: 'I got the paper and I really wanted (to sign up for) the veterinary' // 'because I always wanted to'. The boundary (after VETERINARY) is marked by smaller and less prominent cues: head movement and manual duration (hand hold).

Note: The figure presents the signs at the boundary of a relation but omits the rest of the lexical material due to space limitations.

Source: from Dachkovsky et al. (2022); reproduced with permission.

While the conceptualization of a discourse hierarchy is part of our cognitive capacity from the start of communication, it is the body which is the key to the complex unfolding of linguistic structure over time. The fact that younger ISL signers systematically map their bodily articulators to hierarchical information—using larger articulators at higher levels and smaller, less salient articulators at lower levels of the hierarchy—indicates that the organization of grammar is subject to the passage of time.

8.5 Summary and conclusion

In this chapter, we observe that young sign languages are an empirical source of evidence for language emergence. In addition, we provide evidence that it is the body, the vessel through which signers transmit language, that is a methodological tool for revealing the growing complexities of language as it emerges. Spoken languages are

9780192886491_Book.indb 158 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:39

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

limited in this regard; they are much older or descended from older languages, and we have no record of their origins. Since the link between articulators of the vocal tract and linguistic structure is exceedingly indirect, we can turn to sign languages. Here, we show that sign languages, which can arise anew at any time, and in which the link between physical and linguistic structure is direct, serve as a way to observe language emergence. The methodology we propose is to study the ways in which the body manifests both linguistic structure and its emergence.

Sign languages offer us two important directions for understanding language evolution. First, they show us the fundamental elements necessary for a language to function minimally. Second, they reveal the ordering of the emergence of linguistic properties, the path along which one kind of structure follows, or is dependent on another. When comparing signing used across different generations of a young sign language, the body can more readily point to 'what comes first' (Sandler 2016). Although contemporary sign languages arise in what Arbib (2012) calls 'the language-ready brain', they still offer the only empirical foundation for reconstructing the process of language evolution in humans. We see that the fundamental features of early sign languages include holistic words and prosodic organization of semantically related words (Sandler 2016) followed by structuring of syntax and discourse, not only appearing later, but relying on constituent relations expressed by the body to elaborate relations among structures.

Other researchers, too, suggest that visible bodily actions evolved as part of our communicative endowment (Corballis 2003; Armstrong and Wilcox 2007; Arbib 2012), and evolutionary biologists take the body seriously in understanding language evolution (e.g. Donald 1991; Fitch 2010). The visual-gestural medium of sign languages, in which movements of different articulators (hands, head, face, torso) visually manifest different linguistic functions, enables a form–function mapping more directly than is possible for spoken languages. This mapping, which we call the Grammar of the Body (after Sandler 2012), has become the basis for multiple studies whose focus is the emergence of complexity in young, emerging sign languages. Young sign languages and the method of bodily analysis outlined here offer us a uniquely empirical and plausible reference point for models of language evolution.

Acknowledgements

This work was funded by grants from the Israel Science Foundation, the U.S.-Israel Binational Science Foundation, the U.S. National Institutes of Health, and the European Research Council. We would like to express our deepest appreciation to Mark Aronoff and Carol Padden, and to our late colleague and dear friend Irit Meir.

References

Arbib, M. A. (2012). *How the Brain Got Language: The Mirror System Hypothesis*. New York: Oxford University Press.

9780192886491_Book.indb 159 /12_first_proofs/first_pr…/xml_for_typesetting 27-Mar-25 23:02:39

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Ariel, M. (1991). 'The function of accessibility in a theory of grammar', *Journal of Pragmatics* 16(5): 443–463, https://doi.org/10.1016/0378-2166(91)90136-L.

Armstrong, D. F. and Wilcox, S. E. (2007). *The Gestural Origin of Language*. New York: Oxford University Press, https://doi.org/10.1093/acprof:oso/9780195163483.001.0001.

Baker-Shenk, C. L. and Padden, C. (1978). *American Sign Language: A Look at its History, Structure, and Community*. Silver Spring, MD: T.J. Publishers.

Birkner, K. (2012). 'Prosodic formats of relative clauses in spoken German', in P. Bergmann, J. Brenning, M. Pfeiffer, and E. Reber (eds.), *Prosody and Embodiment in Interactional Grammar* Vol. 18, 19–39. Frieburg: de Gruyter, https://eref.uni-bayreuth.de/id/eprint/20137/.

Corballis, M. C. (2003). *From Hand to Mouth*, https://press.princeton.edu/books/paperback/9780691116730/from-hand-to-mouth.

Cristofaro, S. (2005). *Subordination*. Oxford: Oxford University Press.

Croft, W. (1995). 'Autonomy and functionalist linguistics', *Language* 71(3): 490, https://doi.org/10.2307/416218.

Dachkovsky, S. (2018). *Grammaticalization of Intonation in Israeli Sign Language: From Information Structure to Relative Clause Relations*. [PhD dissertation]. University of Haifa.

Dachkovsky, S. (2022). 'Emergence of a subordinate construction in a sign language: Intonation ploughs the field for morphosyntax', *Glossa: A Journal of General Linguistics* 7(1): Article 1, https://doi.org/10.16995/glossa.9074.

Dachkovsky, S. and Sandler, W. (2009). 'Visual intonation in the prosody of a sign language', *Language and Speech* 52(2–3): 287–314, https://doi.org/10.1177/0023830909103175.

Dachkovsky, S., Healy, C., and Sandler, W. (2013). 'Visual intonation in two sign languages*', *Phonology* 30(2): 211–252, https://doi.org/10.1017/S0952675713000122.

Dachkovsky, S., Stamp, R., and Sandler, W. (2022). 'Mapping the body to the discourse hierarchy in sign language emergence', *Language and Cognition* 15(1): 1–33, https://doi.org/10.1017/langcog.2022.25.

de Boer, B., Sandler, W., and Kirby, S. (2012). 'New perspectives on duality of patterning: Introduction to the special issue', *Language and Cognition* 4(4): 251–259, https://doi.org/10.1515/langcog-2012-0014.

Donald, M. (1991). *Origins of the Modern Mind: Three Stages in the Evolution of Culture and Cognition*. Cambridge, MA: Harvard University Press.

Ergin, R., Meir, I., Ilkbaşaran, D., et al. (2018). 'The development of argument structure in Central Taurus Sign Language', *Sign Language Studies* 18(4): 612–639.

Fenlon, J., Denmark, T., Campbell, R., and Woll, B. (2007). 'Seeing sentence boundaries', *Sign Language & Linguistics* 10(2): 177–200, https://doi.org/10.1075/sll.10.2.06fen.

Fitch, W. T. (2010). *The Evolution of Language*. Cambridge: Cambridge University Press, https://doi.org/10.1017/CBO9780511817779.

Fox, B. A. and Thompson, S. A. (1990). 'A discourse explanation of the grammar of relative clauses in English conversation', *Language* 66(2): 297–316.

Givón, T. (2015). 'The diachronic genesis of synchronic syntax', in B. MacWhinney and W. O'Grady (eds.), *The Handbook of Language Emergence*, 201–214. New York: Wiley-Blackwell.

Grosz, B. J. and Sidner, C. L. (1986). 'Attention, intentions, and the structure of discourse', *Computational Linguistics* 12(3): 175–204.

Hobbs, J. R. (1979). 'Coherence and coreference*', *Cognitive Science* 3(1): 67–90, https://doi.org/10.1207/s15516709cog0301_4.

Kisch, S. (2004). 'Negotiating (genetic) deafness in a Bedouin community', in J. V. van Cleve (ed.), *Genetics, Disability and Deafness*, 148–173. Washington, DC: Gallaudet University Press.

9780192886491_Book.indb 160 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:39

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Labov, W. (1963). 'The social motivation of a sound change', *WORD* 19(3): 273–309, https://doi.org/10.1080/00437956.1963.11659799.

Labov, W. (1966). *The Social Stratification of English in New York City* (2nd ed.). Cambridge: Cambridge University Press, https://doi.org/10.1017/CBO9780511618208.

Lenneberg, E. H. (1967). *Biological Foundations of Language*. New York: Wiley.

Liebal, K., Vaish, A., Haun, D., and Tomasello, M. (2014). 'Does sympathy motivate prosocial behaviour in great apes?', *PLOS ONE* 9(1): e84299, https://doi.org/10.1371/journal.pone.0084299.

Mann, W. and Thompson, S. (1988). *Rhetorical Structure Theory: Toward a Functional Theory of Text Organization*, https://doi.org/10.1515/text.1.1988.8.3.243.

Mayer, J., Jasinskaja, E., and Kölsch, U. (2006). 'Pitch range and pause duration as markers of discourse hierarchy: Perception experiments (Vol. 1)'. INTERSPEECH 2006—ICSLP, *Ninth International Conference on Spoken Language Processing*, Pittsburgh, PA, 17–21 September.

McWhorter, J. (2022). 'Spoken and sign language emergence: A comparison', *Languages* 7(3): Article 3, https://doi.org/10.3390/languages7030184.

Meir, I. and Sandler, W. (2007). *A Language in Space: The Story of Israeli Sign Language* (1st edition). New York: Psychology Press.

Meir, I. and Sandler, W. (2019). 'Variation and conventionalization in language emergence', in E. Doron, M. Rappaport, Y. R. Hovav, and M. Taube (eds.), *Language Contact, Continuity and Change in the Genesis of Modern Hebrew*, 337–363. Amsterdam, Philadelphia, PA: John Benjamins Publishing Company, https://benjamins.com/catalog/la.256.13mei.

Meir, I., Padden, C., Aronoff, M., and Sandler, W. (2013). 'Competing iconicities in the structure of languages', *Cognitive Linguistics* 24(2), https://doi.org/10.1515/cog-2013-0010.

Meir, I., Aronoff, M., Börstell, C., et al. (2017). 'The effect of being human and the basis of grammatical word order: Insights from novel communication systems and young sign languages', *Cognition* 158: 189–207, https://doi.org/10.1016/j.cognition.2016.10.011.

Mithun, M. (2009). 'Adding intonation: Re(e)volving complexity', in T. Givón and M. Shibatani (eds.), *Syntactic Complexity: Diachrony, Acquisition, Neuro-cognition, Evolution*, 53–80. Amsterdam, Philadelphia, PA: John Benjamins Publishing Company, https://doi.org/10.1075/tsl.85.03ree.

Nespor, M. and Sandler, W. (1999). 'Prosody in Israeli Sign Language', *Language and Speech* 42(2–3): 143–176, https://doi.org/10.1177/00238309990420020201.

Nespor, M. and Vogel, I. (1986). 'Prosodic phonology', *Phonology* 5(1): 161–168, https://doi.org/10.1017/S0952675700002219.

Oña, L. S., Sandler, W., and Liebal, K. (2019). 'A stepping stone to compositionality in chimpanzee communication', *PeerJ* 7: e7623, https://doi.org/10.7717/peerj.7623.

Padden, C., Meir, I., Aronoff, M., and Sandler, W. (2010). 'The grammar of space in two new sign languages', in D. Brentari (ed.), *Sign Languages: A Cambridge Survey*, 573–595. Cambridge: Cambridge University Press.

Pfau, R., Steinbach, M., and Woll, B. (eds.). (2012). *Sign Language: An International Handbook*. Berlin: De Gruyter, https://doi.org/10.1515/9783110261325.

Polanyi, L. (1988). 'A formal model of the structure of discourse', *Journal of Pragmatics* 12(5): 601–638, https://doi.org/10.1016/0378-2166(88)90050-1.

Sandler, W. (1999). 'The medium and the message: Prosodic interpretation of linguistic content in sign language', *Sign Language and Linguistics* 2(2): 187–216.

Sandler, W. (2010). 'Prosody and syntax in sign languages', *Transactions of the Philological Society* 108(3): 298–328, https://doi.org/10.1111/j.1467-968X.2010.01242.x.

9780192886491_Book.indb 161 /12_first_proofs/first_pr[illegible]/xml_for_typesetting 27-Mar-25 23:02:39

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Sandler, W. (2012). 'Dedicated gestures and the emergence of sign language', *Gesture* 12(3): 265–307.

Sandler, W. (2013). 'Vive la différence: Sign language and spoken language in language evolution', *Language and Cognition* 5(2–3): 189–203, https://doi.org/10.1515/langcog-2013-0013.

Sandler, W. (2016). 'What comes first in language emergence?', in N. Enfield (ed.), *Dependencies in Language: On the Causal Ontology of Language Systems. Studies in Diversity in Linguistics 99*: 67–86. Berlin: Language Science Press.

Sandler, W. (2017). 'The challenge of sign language phonology', *Annual Review of Linguistics* 3(1): 43–63, https://doi.org/10.1146/annurev-linguistics-011516-034122.

Sandler, W. (2018). 'The body as evidence for the nature of language', *Frontiers in Psychology* 9: 1782, https://doi.org/10.3389/fpsyg.2018.01782.

Sandler, W. and Lillo-Martin, D. (2006). *Sign Language and Linguistic Universals*. Cambridge: Cambridge University Press.

Sandler, W., Meir, I., Padden, C., and Aronoff, M. (2005). 'The emergence of grammar: Systematic structure in a new language', *Proceedings of the National Academy of Sciences* 102(7): 2661–2665, https://doi.org/10.1073/pnas.0405448102.

Sandler, W., Aronoff, M., Meir, I., and Padden, C. (2011). 'The gradual emergence of phonological form in a new language', *Natural Language & Linguistic Theory* 29(2): 503–543, https://doi.org/10.1007/s11049-011-9128-2.

Sankoff, G. and Brown, P. (1976). 'The origins of syntax in discourse: A case study of Tok Pisin relatives', *Language* 52(3): 631–666.

Scott, D. A., Carmi, R., Elbedour, K., et al. (1996). 'An autosomal recessive nonsyndromic-hearing-loss locus identified by DNA pooling using two inbred Bedouin kindreds', *American Journal of Human Genetics* 59(2): 385–391.

Selkirk, E. (1984). *Phonology and Syntax: The Relation Between Sound and Structure*. Cambridge, MA, London: The MIT Press.

Stokoe, W. (1960). 'Sign language structure: An outline of the visual communication systems of the American deaf'. *Journal of Deaf Studies and Deaf Education* 10(1): 3–37.

Van Kuppevelt, J. (1995). 'Discourse structure, topicality and questioning', *Journal of Linguistics* 31(1): 109–147, https://doi.org/10.1017/S002222670000058X.

Wilbur, R. B. (1994). 'Eyeblinks & ASL phrase structure', *Sign Language Studies* 84: 221–240.

9780192886491_Book.indb 162 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:39

CHAPTER 9

SILENT GESTURE

Gesture studies with hearing participants

DANIELLE NAEGELI AND MARIEKE SCHOUWSTRA

Language evolution experiments are an important way to investigate biases and mechanisms that have shaped the languages we know today. Silent gesture experiments are a branch of experiments in which hearing participants (with no knowledge of existing sign languages) improvise to convey information using their hands and bodies. This way, we can simulate, in a controlled environment, situations in which new language systems emerge 'from scratch'. In this chapter, we will describe the methodology and the ways in which it has been applied.

Lab experiments like artificial language learning experiments, or communication games, offer the possibility to observe the evolution of structure and conventions on a small scale in an environment where different pressures can be applied in a controlled manner. However, participants in these experiments are all experienced language users, and therefore not naive to linguistic structure; they might be biased towards the structure used in the language they speak, and this bias has to be considered when analysing any results. One way of alleviating this problem is to conduct experiments with participants of many different language backgrounds; another option is to reduce the effect of a bias towards known linguistic structure through a change in modality.

Most participants in language evolution studies use spoken languages and are not experienced with languages in the visual–gestural modality like sign languages. One way to make use of this inexperience is to design experiments employing a paradigm often called *silent gesture*. In this paradigm, hearing participants, who have no experience with sign languages, are asked to perform tasks in which they need to communicate information using only their hands (and bodies) instead of speech, similarly to the party game *charades*. Silent gesture(s) are different from co-speech gestures. People perform the latter often unconsciously, alongside their speech. Silent gestures, by contrast, do not *accompany* other communicative output, but carry the full communicative burden (Goldin-Meadow and Brentari 2017). We use the term *silent gesture*, however there is also literature using a similar experimental paradigm under the

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

name *pantomime*. In pantomime experiments, participants use their whole bodies, whereas in silent gesture experiments the focus is generally placed on, but not limited to, participants' non-verbal communication expressed torso-upwards (see Żywiczyński et al. 2018, Marentette et al. 2020 for more detailed characterization of silent gesture and pantomime).

Goldin-Meadow et al. (2008) provide a seminal example of how silent gesture can give insight into linguistic behaviour that is not influenced by structures participants are familiar with. In the experiment, participants were asked to describe transitive motion events they saw on vignettes, e.g. 'woman turns knob' using only gestures while being recorded. In the vignettes, all three components were presented simultaneously, however, when reproducing the information participants were forced to choose an order to follow—akin to word order present in many languages. Goldin-Meadow et al. (2008) conducted their experiment with participants from four different language backgrounds, English, Chinese, Spanish, and Turkish. Although Subject-Verb-Object is the dominant word order for most of the participants' spoken languages (English, Chinese, and Spanish), in the experiment, participants across all language groups chose to order the information in Subject-Object-Verb order. The authors conclude that this might be evidence for the existence of a natural order of events, independent from structures in people's spoken languages.

In Goldin-Meadow et al. (2008), participants performed their task alone to a camera after being shown the stimulus vignettes. This kind of on-the-spot improvisation may provide access to cognitive biases in individuals. Cognitive biases, however, are not the only relevant force in shaping conventional structures, since languages do not emerge in isolation. Instead, they are used for communication between individuals, and have been passed on to generations of new language users. These mechanisms have also been simulated in the lab, in experiments in which silent gesture is combined with learning and interaction.

It is possible to compare silent gesture lab results to existing languages. Many papers relate their results to typological patterns in spoken (e.g. Culbertson et al. 2020) and signed languages (e.g. Slonimska et al. 2022). A link that has also been explored is that between silent gesture and structures observed in young sign languages. Examples of young sign languages are Kata Kolok, Al-Sayyid Bedouin Sign Language, and Nicaraguan Sign Language (see Chapter 8, this volume on emerging sign language); they emerged in situations where deaf individuals did not share a conventionalized system. When comparing emerging linguistic structures in the lab to young sign languages, it is important to keep in mind that the situation in which participants produce silent gestures in the lab is vastly different from any natural language emergence scenario. In lab experiments, the meaning space is often purposefully constrained and controlled and thus, findings in silent gesture cannot claim to inform us directly on how natural sign languages evolved. Nonetheless, looking to findings from young sign languages and the communities using them can inform laboratory experiments on what features or pressures to isolate and to investigate. Moreover, they can supply a frame of reference for experimental results.

9780192886491_Book.indb 164 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:39

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

As pointed out above, silent gesture experiments can be run with multiple participants to investigate the potential influence of communicative interaction on emerging language structure, for instance in a director–matcher task. In such a task, participants are assigned roles as director, conveying information, and matcher, trying to match this information to different options in an array. In the next part of the chapter, we will give a more in-depth account of a recent paper published by us that uses both a single-participant and a dyadic setup to further investigate word order in silent gesture. Subsequently, we will systematically discuss methodological choices that need to be made when working with this paradigm.

9.1 Focus paper: Reversibility and regularity in silent gesture production and interaction

In what follows, we will give a detailed description of the two experiments in Schouwstra et al. (2022). This will allow us to demonstrate how the silent gesture paradigm can be used. We will focus on central questions targeted with this work, followed by the details of the setup, the stimuli, and the data analysis. The study builds on the findings of Goldin-Meadow et al. (2008) and subsequent research into word order conventions in silent gesture and places its focus on reversible events. Reversible events are generally defined as events in which both agent and patient are animate as in 'girl kicks boy' and therefore their roles could be reversed. This is contrasted with non-reversible events, in which the patient is inanimate, like in 'girl kicks ball'. In these events, a reversal of the roles is unlikely. While Goldin-Meadow et al. (2008) found that participants with different language backgrounds prefer to order transitive events in an SOV order, all of the transitive events used in that study were non-reversible. More recent work targeting reversible events has shown that participants prefer SVO order for these events (e.g. Gibson et al. 2013; Hall et al. 2013; Futrell et al. 2015; Meir et al. 2017). Two prominent explanations for this pattern have been put forward. One of them, the *noisy channel account*, is based in information theory, which understands communication to be inherently noisy: language users cannot assume that messages are transferred to others without loss of information, and therefore optimize the structures they use to minimize the chance of error (Gibson et al. 2013; Futrell et al. 2015). The noisy channel account attributes the preference for SVO ordering of reversible events to this aim. Even if the beginning or the end of the communication is unclear or lost, as the verb is in the middle position the recipient could infer if the argument successfully conveyed is agent or patient depending on its position relative to the verb.[1] Another approach, the *role conflict account* (Hall et al.

[1] For this to be applicable, an additional assumption, a subject-first preference, is needed. Evidence for this is provided by Futrell et al. (2015). Evidence of a closely related preference, human-first, is provided in Meir et al. (2017).

9780192886491_Book.indb 165 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:39

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

2013) argues that the preference of SVO order for reversible events is entirely modality specific. Producers of silent gesture tend to convey information about the action while embodying the role of the subject ('body as subject'; Meir et al. 2007). This makes SOV order inconvenient as gesturers would be forced to switch from embodying the role of subject to embodying the role of direct object and back to the role of subject 'performing' the action. When using an SVO order for a reversible event this issue is resolved.

These two accounts explain the change in word order preference as resulting from different pressures, the former linking it to successful communication while the latter account links the different word order preference directly to modality and production-specific pressures. In Schouwstra et al. (2022), we investigate these two explanations by first analysing the nature of reversible events. When discussing the reversibility of an event the spotlight is usually placed (solely) on whether the direct object is animate. However, we argue that the potential reversibility of an event is a composite phenomenon that does not only depend on the animacy of the object but also on the properties of the verb in the event. A verb like 'to kick' can quite feasibly be used in a reversible event ('girl kicks boy') as well as a non-reversible event ('girl kicks ball'). Yet, for a verb like 'to eat', while it is technically grammatically correct to reverse an event like 'the man eats a sandwich' into 'the sandwich eats a man' it is far less likely to occur in that order. Thus, events cannot just be separated into reversible and non-reversible, but the category of non-reversible events can be split into two subcategories. The events with actions that are as likely to co-occur with an animate as with an inanimate object we call weakly non-reversible events (e.g. 'boy kicks box'). By contrast, the events which are unlikely to be reversed due to the properties of the action (e.g. 'boy eats pizza') we named strongly non-reversible events.

In our study, we set out to investigate this more fine-grained account of reversibility, hypothesizing that we should observe a difference in word order preference between weakly and strongly non-reversible events. To do this, we used stimuli of all three event types, reversible, weakly non-reversible, and strongly non-reversible (see Figure 9.1 for example stimulus pictures), in two different experimental setups, a single-participant setup and an interactive setup.



FIGURE 9.1 Examples of stimulus pictures for the different event types.

'Clown cuts hat' for strongly non-reversible, 'pirate punches vase' for weakly non-reversible, and 'burglar points at nun' for reversible events. The crucial difference between the first two images is in the action: 'punch' can more plausibly be used with an animate patient than 'cut', which makes the 'cut' event more strongly non-reversible than the 'punch' event.

9780192886491_Book.indb 166 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:40

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

For the first experiment, participants were asked to improvise gestures to communicate events in a single-participant setting while being recorded. They described stimulus pictures, presented to them one at a time on a tablet to a camera.[2] The stimulus pictures consisted of twelve line drawings per event type resulting in a total of thirty-six pictures.[3] None of the actions used in the drawings overlapped across event types. Participants were assigned randomly to one of two conditions which differed in whether they saw the block of reversible events or non-reversible events (both weakly and strongly non-reversible events) first.[4] Within the two blocks of event types, the stimuli were shown in a random order and each condition had two practice items (of the same event type as the events in their first block) before the experiment started. We included these training trials to ensure participants understood the task and to give them a chance to get accustomed to communicating with only gestures. Participants (fifteen per condition) all had English as their first language and no previous experience with sign languages. The experimenter was present in the lab with the participant sitting across the table. For each trial, once the participant had finished their gesture string the experimenter triggered the next stimulus presentation. The video recordings were hand-coded for word order and for the analysis we categorized them as SOV, SVO, and Other.

After analysing the proportion of SOV order use according to condition and event type,[5] we observed that the type of verb used in the event did have a significant effect on the word order used in silent gesture. Participants used SOV order to a higher degree when communicating strongly non-reversible events than when communicating weakly non-reversible events. Thus, the results of the first experiment in this study support a more complex notion of reversibility as a composite. It also provides evidence that the role conflict account cannot be the sole explanation for the different word order preference in silent gesture across different events. Both strongly and weakly non-reversible events comprise of an animate human agent and an inanimate patient and yet they elicit different ordering behaviours.[6] The noisy channel account, on the other hand, *is* compatible with our results: weakly non-reversible events are slightly more susceptible to noise than strongly non-reversible events, and SVO is the more noise-resistant order according to this account.

In our second experiment, we focused on word order consistency and with that the emergence of a word order convention. By consistently using one word order, communicators can effectively eliminate confusion regarding who is the agent and the

[2] A Logitec wide-angle lens webcam.

[3] All stimuli are accessible on https://osf.io/unw98.

[4] We opted for this stimulus presentation instead of randomly mixing event types to control for the possibility the mere inclusion of reversible events in combination with participants feeling the need to produce consistent output might push for a SVO order dominance.

[5] We used a binomial mixed-effects regression; data and analysis code are accessible on https://osf.io/2pr48/.

[6] We also analysed whether the order of the stimuli blocks that participants were presented with influenced their behaviour but found no evidence for that.

9780192886491_Book.indb 167 /12_first_proofs/first_pr…/xml_for_typesetting 27-Mar-25 23:02:40

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

patient. To investigate the pressure of consistency, the second experiment of our study required participants to communicate with each other using only silent gesture in a dyadic setup (similar setups can be found in work by Christensen et al. 2016; Fehér et al. 2016; and Schouwstra et al. 2020). Two participants were paired up at random and asked to sit at a table across from each other with the experimenter monitoring the experiment from the side. In the experiment, participants were asked to complete a director–matcher task together. In each trial, one participant took on the role of the director and conveyed information about the stimulus to the other participant. The other participant, the matcher, was shown an array of stimuli, consisting of the target picture (the one shown also to the director) and distractors. The matcher's task was to choose the target picture based on the description they received from the director. The participants switched roles each trial. The matcher saw an array of eight stimulus pictures, together showing all eight possible combinations of one of two agents, one of two patients, and one of two actions. In total, we used eight such sets.[7]

Participants were shown the stimulus pictures on a tablet placed in front of them and two webcams were installed in the middle of the table recording each participant independently. In total, participants played through six rounds of thirty-two trials. Every stimulus picture (a total of sixty-four) was the target stimulus at least once and the target stimuli were balanced across event types for every participant. After each trial, both director and matcher received full feedback, i.e. whether the matcher's choice was correct or incorrect and if the latter was the case, the director was also shown the stimulus picture chosen by the matcher while the matcher was also shown the target stimulus on the tablet in front of them. In order to gamify the experiment procedure and engage participants, we introduced a time pressure. Participants were told that their goal was to complete each round of thirty-two trials faster than the previous round. If achieved, the experimenter would reward them with a sweet and, furthermore, participants were made aware that the overall fastest dyad would receive an extra financial reward.

We were able to include the data of twenty-four participants in our analysis. The videos of the participants' silent gesture descriptions were coded for word order and subsequently categorized as SOV, SVO, or Other for the statistical analysis. We used entropy measures[8] for each dyad to test if participants' word order use became more consistent (more predictable) over time. As entropy is a measure of uncertainty, variation in a system can be quantified in this notion (Ferdinand et al. 2019). Increasingly consistent word order usage by participants can be detected by a drop in entropy over the course of the experiment.[9] Our analysis revealed that round number had a significant effect on entropy and, therefore, showed that dyads became more consistent in their word order use. Overall, participants used SVO order the most, followed by Other and SOV, with a drastic increase in SVO order use over time.

[7] All stimuli are accessible on https://osf.io/unw98.

[8] To calculate this, we used Shannon's entropy. A detailed description of this procedure can be found in our paper (Schouwstra et al. 2022).

[9] For all analyses mentioned, data and code are accessible on https://osf.io/2pr48/.

9780192886491_Book.indb 168 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:40

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Focusing on the fine-grained semantic properties of the events, we found that SOV was used significantly more often for strongly non-reversible events than for weakly non-reversible events. Moreover, SOV order use decreased more steeply over time for weakly non-reversible than for strongly non-reversible events. Thus, we found that participants' behaviours differed for the two non-reversible event types and that participants used SVO more and more consistently over the course of the experiment.

To summarize, in our study, we used the silent gesture paradigm to investigate and subsequently redefine the concept of reversibility as not only dictated by the animacy of the patient but also the properties of the action. Further, in a second experiment, we used the silent gesture paradigm in a communicative setting. This way, we showed that participants decreased variation in their output over time opting for more regular word order use. The fact that SVO became by far the most frequent word order in our experiment could potentially be a first language influence after all: our participants were all speakers of English. At the same time, however, SVO is the word order that has been predicted (Futrell et al. 2015) to be the most noise-resistant, particularly for reversible events. In this case, further research with speakers of languages with different dominant orders could help to disentangle the two possible explanations.

Our focus paper showcases how silent gesture can be run as a production task with single participants, but also in a director–matcher setup. In what follows, we will describe the dimensions along which work in this area has varied. Together, this will form an overview of the possible uses of silent gesture to investigate individual biases and cultural mechanisms involved in language emergence.

9.2 Research questions and experimental setups

The study described above focused on basic word order. Other work used the paradigm to investigate the order of noun, adjective, numeral, and demonstrative (Culbertson et al. 2020). But word order is not the only aspect of language that can be studied in silent gesture experiments. Ortega and Özyürek (2020) elicited single words to study the encoding of lexical items, and various iconic strategies, in silent gesture. Iconicity in gesture production is a comprehensive topic in itself and a review of different conceptions can be found in Chapter 10, this volume. Sato et al. (2020) also investigated single words, but focused instead on phonological properties of the gestures to assess how iconicity affects learning of complex forms. Finally, Motamedi et al. (2019) explored the emergence of systematicity in silent gesture, using a meaning space of single words that share functional or thematic properties.

The paper by Motamedi et al. (2019) also demonstrated how silent gesture is versatile and can be implemented in different setups. The authors simulated mechanisms of cultural evolution of language (e.g. Kirby et al. 2015, but see Tamariz and Papa, Chapter 1,

9780192886491_Book.indb 169 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:40

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

this volume), by using silent gesture in an interactive task and an iterated learning setup. The interactive setup simulated language use by asking paired-up participants to communicate information successfully in a director–matcher task. The iterated learning design simulated generational transmission of language by using the output of one participant as training input for the next participant in a chain-like structure. In a third condition, the researchers combined the interactive and the iterated learning setup in order to systematically compare the combined effect of learning and interaction to each effect individually.

Apart from the manipulations described above, details of the experimental setup can potentially influence the resulting structures significantly. However, there are currently not enough experimental studies that investigate those details systematically. Here, we point out the main aspects of silent gesture experiments that have been implemented differently by researchers; the references we provide are examples and not exhaustive.

In single-participant experiments, the researchers can be present in the room (e.g. Schouwstra et al. 2022) or control the experiment software and recording from a different room. While the latter might minimize an observer effect it should be noted that, for example, Kirton et al. (2021) found that when participants were alone in a lab booth, their output showed a high proportion of omitted elements. Apart from that, in interactive silent gesture experiments, the two participants can be present in the same room (Schouwstra et al. 2022) or in different rooms, and communicate through a video connection (Bohn et al. 2019; Motamedi et al. 2021). Conducting an experiment through a video connection means that there is no opportunity for repair strategies between participants. A silent gesture experiment that focused specifically on repair (Micklos 2016) indicates that such strategies can be a crucial process for negotiating a communication system. Also, the details of interaction have been varied: in communicative tasks, some researchers switch the roles of director and matcher each trial, while others do not, and have one participant as the director for many trials in a row. Christensen et al. (2016) applied both approaches in a series of experiments, finding that the change in word order over time was greater when participants switched roles each trial.

Another setup specification to be considered is whether participants are seated or standing. As most hearing people are not very familiar with using only their hands to communicate, this can cause some discomfort and result in participants standing up to express certain concepts with their whole body. While it is dependent on the research at hand whether participants should be seated, remain seated throughout, or be standing from the start, it will affect the camera angle(s) at which the experiment is recorded.

While the silent gesture paradigm lends itself to investigate behaviours in participants' production, more recent experiments have also used silent gesture in perception and learning experiments (Motamedi et al. 2022; Holtz et al. 2022; Do et al. 2022). Motamedi et al. (2022) used pre-recorded silent gesture videos in an online forced-choice setting, and replicated SVO and SOV order preference for intensional events and extensional events respectively, previously demonstrated in a silent gesture production

9780192886491_Book.indb 170 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:40

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

experiment (Schouwstra and de Swart 2014).[10] Participants were shown line drawings of intensional (e.g. 'nun thinks about ukulele') or extensional events (e.g. 'nun throws ukulele') accompanied by two videos with silent gesture descriptions differing solely in gesture constituent order. Participants' task was to click on the video which they felt best described the event in the picture. Moreover, Motamedi et al. (2022) further employed the same pre-recorded silent gesture videos in an artificial language learning experiment. This learning experiment employed four training languages with different word order distributions (different proportions of SOV and SVO). Participants were trained on one of these languages, and then tested in a forced-choice task, again by clicking on the video of their choice. Later work used the same online forced-choice task with pre-filmed silent gesture videos to assess ordering preferences for genitives and adjectives (Holtz et al. 2022) and to compare order preferences in the noun phrase and the verb phrase (Do et al. 2022).

9.3 Stimuli

As we have detailed so far, silent gesture can be elicited with many different types of stimuli: pictures, animations, videos, or written text. It is important to keep in mind that all stimulus types come with their own caveats. For example, pictures (drawings or photos), videos, and animated vignettes offer options that are not tied to linguistic conventions, whereas written text obviously is. In videos, there is an inherent order of how the event unfolds, which might pose a problem as it could affect participants' ordering preferences, if that is the research objective. Images raise a related potential problem as the directional order of components in the picture might influence gesture order. This can be controlled for by using the original picture as well as its mirror image.

When using visual stimuli, the way in which agents and patients are depicted should also be considered as it might influence participants' output. Kirton et al. (2021) compared the saliency of different agents by testing participants' ordering behaviour for the same events with generic and character agents. Generic agents are human agents identifiable by gender and minimal physical characteristics, whereas character agents are composed of agents whose portrayal was linked to their profession or commonly known characteristics (e.g. a pirate). They found that non-salient generic agents were more likely to be entirely omitted by participants than the more salient character agents. Visual stimuli, as well as any other stimuli, should be adjusted according to cultural validity. The depiction of a character agent can be culturally dependent and may vary or not be transparent, e.g. a judge depicted with a powdered wig might be easily recognized as a judge in one cultural setting but not in another. Furthermore, in reversible events

[10] 'Extensional' refers to events where a direct object is somehow manipulated (e.g. 'throwing a guitar') whereas 'intensional' events are events in which the direct object has to be interpreted in relation to the event itself and does not necessarily need to exist independently from it (e.g. 'knitting a scarf').

9780192886491_Book.indb 171 /12_first_proofs/first_pr/xml_for_typesetting 27-Mar-25 23:02:40

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

there might be levels of 'acceptedness' or 'likelihood' of reversible scenarios occurring that differ according to cultures depending on seniority or gender of the actor and patient, e.g. a woman brushing a child's hair would be very likely, but a child brushing a woman's hair, though perfectly possible, would be less likely.

Lastly, Motamedi et al. (2021) used orthographically presented sentences as stimuli in their study on the evolution of argument distinction strategies in a silent gesture experiment. While this type of stimuli does come with linguistic conventions, for their experiment, it was not an issue as they were investigating (spatial) strategies participants developed to differentiate between two possible agents. Rather, for this type of research using sentences, combined with asking participants to refrain from any type of spelling, is preferable to having pictures where agents might be identified, and therefore distinguished, by physical features depicted on them.

9.4 Participants

While being versatile in both setups and stimuli, the silent gesture paradigm has also been successfully conducted with adults and children. Bohn et al. (2019) used a silent gesture-like experimental design in a communication game conducted with 3-, 4-, and 6-year old children. They found that children were able to innovate gestures, with only the youngest group needing continuous prompts. Currently, there is no data that compares adult and children's production directly, e.g. concerning their word order usage.

As touched upon earlier, researchers such as Goldin-Meadow et al. (2008) and many others, have used silent gesture experiments across participant groups of different language backgrounds and showed similarities, seemingly independent from known linguistic structure. Running silent gesture experiments cross-culturally, however, also offers an opportunity to investigate domains where typological differences are especially prevalent. In a methodological replication of Motamedi et al. (2021), Naegeli et al. (2022) used the paradigm to assess grammatical use of space for referential strategies by Balinese and Dutch participants. The two groups differ in their frame of reference and use of space in their co-speech gesture systems, such as employing a geocentric vs. an egocentric pointing system (Wassmann and Dasen 2006). Data collection is still ongoing, but finding strategy differences between the participant groups might indicate that differences in typology are rooted in more domains than just linguistic structure.

9.5 Analyses

Most silent gesture experiments will result in video data. These videos can be analysed both qualitatively and quantitatively. This often requires the formulation of a coding

9780192886491_Book.indb 172 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:40

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

scheme on the basis of which all the relevant video data is coded by hand. The coding scheme should come with clear instructions as at least part of the video data should be second-coded by a different researcher so that coding agreement can be measured. What is coded for depends on the research question, but it could include: binomial measures on whether, for a part of the stimuli, a gesture is present, plain coding of order in which the components of the gesture string appear, location in space in which the gesture is performed,[11] the type of gesture used by the participant,[12] or handedness for each gesture.

A few studies have approached the coding of silent gesture data in a more automated way by tracking gesturers using Microsoft Kinect (Namboodiripad et al. 2016). In the study, participants partaking in a dyadic communication task using silent gesture were recorded by a Microsoft Kinect, a camera and body-tracking sensor which uses infrared light projections to create a depth field. This way the researchers could track the size of the articulatory space and calculate the kinetic energy participants used for their gestures. Moreover, Pouw et al. (2021) revisited the data set from Motamedi et al. (2019) using computer vision techniques to assess whether the emergence of systematicity found in the original analysis is reflected in kinematic patterns (gesture forms).

9.6 Discussion and outlook

In this chapter, we have introduced the silent gesture paradigm. We have highlighted its advantage of changing modality forcing participants to communicate in a novel way and thereby circumventing existing linguistic conventions. We detailed its multipurpose use for investigating cognitive biases and the evolution of language features. In the section on the focus paper, we took an in-depth look at two experiments in which we used the paradigm to investigate the concept of reversibility and the emergence of word order conventions through repeated interaction. Furthermore, we have discussed the paradigm's versatility when employed in different experimental setups. As discussed above, it is suitable for explorations of many research questions when used in production, selection, and learning tasks in single-participant or interactive setups. We have also provided examples of different stimulus options and analysis methods. Silent gesture is a relatively new methodology. Some of the setup variations have not been investigated systematically yet, so we do not always know the consequences of design decisions. What is clear from the work described in this chapter is that silent gesture can be a powerful way to measure how individuals convey information when there is not a shared system of conventions. Extending the paradigm to investigate the influence of learning and various aspects of interaction, and comparison of silent gesture

[11] In neutral or non-neutral space, but also, on the participant's left, right, or in the centre.

[12] E.g. tracing an outline, handling an object, using part of the body to represent the object (for a detailed description of different gesture types refer to Chapter 10, this volume).

9780192886491_Book.indb 173 /12_first_proofs/first_pr…/xml_for_typesetting 27-Mar-25 23:02:40

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

results with patterns in existing languages make it all the more powerful. Finally, silent gesture experiments tend to focus either on phrase/sentence structure or lexical phenomena. For the future, another fruitful area for this paradigm could be its application for investigating how space is conceptualized. Due to its spatial nature and attested distance from learned linguistic conventions, silent gesture can help tease apart cognitive conceptualization of space across cultures from its linguistic implementations.

References

Bohn M., Kachel G., and Tomasello M. (2019). 'Young children spontaneously recreate core properties of language in a new modality', *Proceedings of the National Academy of Sciences USA* 116(51): 26072–26077, https://doi.org/10.1073/pnas.1904871116.

Christensen, P., Fusaroli, R., and Tylén, K. (2016). 'Environmental constraints shaping constituent order in emerging communication systems: Structural iconicity, interactive alignment and conventionalization', *Cognition* 146: 67–80.

Culbertson, J., Schouwstra, M., and Kirby, S. (2020). 'From the world to word order: Deriving biases in noun phrase order from statistical properties of the world', *Language* 96(3): 696–717.

Do, M. L., Kirby, S., and Goldin-Meadow, S. (2022). 'Regularization of word order in the verb phrase differs from the noun phrase: Evidence from an online silent gesture perception paradigm', *Proceedings of the Annual Meeting of the Cognitive Science Society* 44, https://escholarship.org/uc/item/497529dx.

Fehér, O., Wonnacott, E., and Smith, K. (2016). 'Structural priming in artificial languages and the regularisation of unpredictable variation', *Journal of Memory and Language* 91: 158–180.

Ferdinand, V., Kirby, S., and Smith, K. (2019). 'The cognitive roots of regularization in language', *Cognition* 184: 53–68.

Futrell, R., Hickey, T., Lee, A., et al. (2015). 'Crosslinguistic gestures reflect typological universals: A subject-initial, verb-final bias in speakers of diverse languages', *Cognition* 136: 215–221.

Gibson, E., Piantadosi, S. T., Brink, K., et al. (2013). 'A noisy-channel account of crosslinguistic word-order variation', *Psychological Science* 24: 1079–1088, doi: 10.1177/095679.

Goldin-Meadow, S. and Brentari, D. (2017). 'Gesture, sign, & language: The coming of age of sign language and gesture studies', *Behavioral and Brain Sciences* 40: E46, doi:10.1017/S0140525X15001247.

Goldin-Meadow, S., So, W. C., Ozyürek, A., and Mylander, C. (2008). 'The natural order of events: How speakers of different languages represent events nonverbally', *Proceedings of the National Academy of Sciences of the United States of America* 105(27): 9163–9168, https://doi.org/10.1073/pnas.0710060105.

Hall, M. L., Mayberry, R. I., and Ferreira, V. S. (2013). 'Cognitive constraints on constituent order: Evidence from elicited pantomime', *Cognition* 129: 1–17, doi: 10.1016/j.cognition.2013.05.004.

Holtz, A., Kirby, S., and Culbertson, J. (2022). 'The influence of category-specific and system-wide preferences on cross-linguistic word order patterns', *Proceedings of the Annual Meeting of the Cognitive Science Society* 44.

Kirby, S., Tamariz, M., Cornish, H., and Smith, K. (2015). 'Compression and communication in the cultural evolution of linguistic structure', *Cognition* 141: 87–102, https://doi.org/10.1016/j.cognition.2015.03.016.

9780192886491_Book.indb 174 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:40

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Kirton, F., Kirby, S., Smith, K., et al. (2021). 'Constituent order in silent gesture reflects the perspective of the producer', *Journal of Language Evolution* 6: 54–76, doi: 10.1093/jole/lzaa010.

Marentette, P., Furman, R., Suvanto, M., and Nicoladis, E. (2020). 'Pantomime (not silent gesture) in multimodal communication: Evidence from children's narratives', *Frontiers in Psychology* 11: 575952, 10.3389/fpsyg.2020.575952.

Meir, I., Padden, C. A., Aronoff, M., and Sandler, W. (2007). 'Body as subject1', *Journal of Linguistics* 43: 531–563, doi: 10.1017/S0022226707004768.

Meir, I., Aronoff, M., Börstell, C., et al. (2017). 'The effect of being human and the basis of grammatical word order: Insights from novel communication systems and young sign languages', *Cognition* 158: 189–207, doi: 10.1016/j.cognition.2016.10.011.

Micklos, A. (2016). 'Interaction for facilitating conventionalization: Negotiating the silent gesture communication of noun-verb pairs', in S. G. Roberts, C. Cuskley, L. McCrohon, et al. (eds.), *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANG11)*, http://evolang.org/neworleans/papers/143.html.

Motamedi, Y., Schouwstra, M., Smith, K., et al. (2019). 'Evolving artificial sign languages in the lab: From improvised gesture to systematic sign', *Cognition* 192: 103964, https://doi:10.1016/j.cognition.2019.05.001.

Motamedi, Y., Smith, K., Schouwstra, M., et al. (2021). 'The emergence of systematic argument distinctions in artificial sign languages', *Journal of Language Evolution* 6: 77–98, doi: 10.1093/jole/lzab002.

Motamedi, Y., Wolters, L., Naegeli, D., et al. (2022). 'From improvisation to learning: How naturalness and systematicity shape language evolution', *Cognition* 228: 105206, https://doi.org/10.1016/j.cognition.2022.105206.

Naegeli, D., Peeters, D., Krahmer, E., et al. (2022). 'The role of cross-cultural differences in referential strategy emergence: Data from artificial sign language learning', in A. Ravignani, R. Asano, D. Valente, et al. (eds.), *The Evolution of Language: Proceedings of the Joint Conference on Language Evolution* (JCoLE), 539–541. Nijmegen: Joint Conference on Language Evolution (JCoLE), https://doi:10.17617/2.3398549.

Namboodiripad, S., Lenzen, D., Lepic, R., and Verhoef, T. (2016). 'Measuring conventionalization in the manual modality', *Journal of Language Evolution* 1(2): 109–118, https://doi.org/10.1093/jole/lzw005.

Ortega, G. and Özyürek, A. (2020). 'Systematic mappings between semantic categories and types of iconic representations in the manual modality: A normed database of silent gesture', *Behavior Research Methods* 52(1): 51–67.

Pouw, W., Dingemanse, M., Motamedi, Y., and Özyürek, A. (2021). 'A systematic investigation of gesture kinematics in evolving manual languages in the lab', *Cognitive Science* 45(7): e13014, https://doi.org/10.1111/cogs.13014.

Sato, A., Schouwstra, M., Flaherty, M., and Kirby, S. (2020). 'Do all aspects of learning benefit from iconicity? Evidence from motion capture', *Language and Cognition: An Interdisciplinary Journal of Language and Cognitive Science* 12(1): 36–55, https://doi.org/10.1017/langcog.2019.37.

Schouwstra, M. and de Swart, H. (2014). 'The semantic origins of word order', *Cognition* 131: 431–436, doi: 10.1016/j.cognition.2014.03.004.

Schouwstra, M., Smith, K., and Kirby, S. (2020). The emergence of word order conventions: Improvisation, interaction and transmission, https://doi.org/10.31234/osf.io/wdfu2.

9780192886491_Book.indb 175 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:40

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Schouwstra, M., Naegeli, D., and Kirby, S. (2022). 'Investigating word order emergence: Constraints from cognition and communication', *Frontiers in Psychology* 13: 805144, doi: 10.3389/fpsyg.2022.805144.

Slonimska, A., Özyürek, A., and Capirci, O. (2022). 'Simultaneity as an emergent property of efficient communication in language: A comparison of silent gesture and sign language', *Cognitive Science* 46(5): e13133, https://doi.org/10.1111/cogs.13133.

Wassmann, J. and Dasen, P. R. (2006). 'How to orient yourself in Balinese space: Combining ethnographic and psychological methods for the study of cognitive processes', in J. Straub, D. Weidemann, C. Köbl, and B. Zielke (eds.), *Pursuit of Meaning: Advances in Cultural and Cross-Cultural Psychology*, 351–376. Bielefeld: Transcript Publishers.

Żywiczyński, P., Wacewicz, S., and Sibierska, M. (2018). 'Defining pantomime for language evolution research', *Topoi* 37(2): 307–318.

9780192886491_Book.indb 176 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:40

CHAPTER 10

METHODS TO STUDY EVOLUTION OF ICONICITY IN SIGN LANGUAGES

ANITA SLONIMSKA AND ASLI ÖZYÜREK

10.1 Introduction

Sign languages—conventional languages of deaf communities—have been considered to provide a window to answer some questions regarding language emergence and evolution. Reasons for this connection are manyfold, including a recent discovery of the existence of emerging sign languages (see Chapter 8, this volume), which, for example, also offer to study the role of population structure in language evolution (Lutzenberger et al. 2023), among many others. Another reason for this connection, the one we focus on in this chapter, is the prevalence of iconic structures in sign languages—which has led language researchers to speculate that iconicity,[1] i.e. 'existence of a structure-preserving mapping between mental models of linguistic form and meaning' (Taub 2001: 23) can be a starting point (e.g. 'protolanguage', Arbib 2005; Tomasello 2008) for language emergence and language evolution in general.

Sign languages include high proportions of iconic structures at many levels, from lexicon to syntax to discourse. Therefore, this property of sign languages has been considered to reveal some insights into how languages evolve. This is partly because sign languages, even the most established ones—have been around not more than 300 years (Kyle et al. 1988; Sandler 2018) and thus are much younger than any spoken language we know of. Thus, it has been suggested that the structures and properties we see even in established sign languages—can be suggestive of 'earlier stages' in language evolution

[1] In the present chapter, we focus on iconicity in the visual modality. However, note that iconic mapping between linguistic form and meaning is also available in the vocal modality as in the case of onomatopoeias and ideophones (Dingemanse 2011, 2015).

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

in a way we would never be able to witness in spoken languages. More specifically, it has been a recurrent theme to assume that iconic structures in sign languages are precursors of arbitrary and segmental categorical structures found in spoken languages (e.g. Senghas et al. 2004; Aronoff et al. 2005).

However, as Jackendoff (2010: 63) has said, 'your theory of language evolution depends on your theory of language'. Recently, it has been debated as to whether the omnipresence of iconicity in sign languages indeed exhibits some 'early stages' of language evolution or whether it exhibits one of the core aspects of language 'proper' within established sign language structures (e.g. Cuxac and Sallandre 2007; Padden et al. 2013). According to the latter view, one needs to have a more nuanced perspective regarding the role of iconicity in language evolution. This requires first defining iconicity at different levels of linguistic organization and rigorous empirical methods to study it systematically. Therefore, in this chapter, we focus on iconicity and review the literature on how it has been defined and methodologically investigated within established sign languages as well as in relation to its role in language evolution.

Many studies in established sign languages have recently attempted to understand the role of iconicity in sign language grammar and uses. These studies involve methods first to assess and define the type of iconicity and their semiotics in the existing sign language lexicon (i.e. as in sign language databases, dictionaries, etc.) or to elicit signs at the lexical or discourse levels through picture or movie descriptions. Some others use iconicity rating methodology to gain insights into the degree of iconicity of the linguistic signs at the lexical level.

Comparisons are also made between iconic structures in sign languages and iconic structures used by hearing people in *silent gesture* or speech-accompanying gestures on lexical as well as discourse levels. In such comparisons, gesture (silent or co-speech gesture) functions as a 'baseline' to be able to identify the linguistic status of the modality-specific properties such as iconicity at lexical and discourse levels in established sign languages.

Such comparisons are especially useful for understanding language evolution as they have the potential to reveal which structures, which we find in established sign languages, are a result of an evolutionary change and which structures may be afforded by the visual modality from the get-go. Such a perspective opens a door to exploring fundamental questions in language evolution: does iconicity continue to erode over time in favour of abstract and arbitrary forms as the linguistic system matures? Does iconicity undergo an evolutionary change from being a general affordance of the visual modality to a linguistic resource in sign language structuring? And if so, what are the pressures that influence its evolution?

Before one can begin to investigate these questions, it is first necessary to have a clear definition of iconicity as well as a thorough understanding of how, and to what extent, iconicity is used in established sign languages. Thus, in this chapter, we will outline such different approaches and methods to iconicity and how one can operationalize them on lexical and discourse levels (corresponding to different types of iconicity) to investigate its role in sign language structures and, ultimately, its role in language evolution.

9780192886491_Book.indb 178 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:40

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

In the chapter, we will also introduce recent kinematic measurements (e.g. the form of the signs) as approaches to studying movement patterns in the visual modality and what they might reveal about the role of iconicity in the evolution of sign language structures.

We start by focusing on iconicity at the lexical level by providing a comprehensive definition and the ways it can be categorized based on the semiotics of the mapping between the sign and its meaning (section 10.2.1). We then turn to describe two methods that have been used to investigate iconicity in lexicons of sign languages and their potential applications in language evolution research (sections 10.2.2 and 10.2.3). We then shift our focus to a method that allows studying the evolution of iconicity used at higher-lever structures beyond the lexical level which uses simultaneous (as opposed to sequential structures) organization of linguistic information (section 10.3).

10.2 Methods to study the evolution of iconicity at the lexical level

10.2.1 Iconicity in sign languages at the lexical level

Iconicity in sign languages is generally defined as, 'a structure-preserving mapping between mental models of linguistic form and meaning' (Taub 2001: 23). On the lexical level, this implies that the linguistic form bears some sort of a perceptual resemblance to the concept it is denoting. In sign languages, visual modality allows for a variety of visually motivated mappings between the concept and the sign.

Some researchers view iconicity as a feature that allows connecting human experience to linguistic form through direct mapping (Thompson et al. 2012; Perniss and Vigliocco 2014; Vinson et al. 2015; Sidhu et al. 2020). Others have argued that iconicity is better viewed as a structural mapping between a linguistic form and a mental representation in which human experience is schematized rather than being a direct link to the experience itself (Taub 2001; Emmorey 2014). More specifically, according to the *analogue-building model*, lexical signs are constructed in a multistep process (Taub 2001; Emmorey 2014): image representing the concept selection → schematization of the selected image → mapping schematized image onto the linguistic form. For example, a sign for a SPOON in Turkish Sign Language (TID) is represented by a one-handed sign in which the concept is schematized by selecting a specific handshape representing the holding of the spoon and movement towards the mouth (Figure 10.1a). In turn, in Czech Sign Language (ČZJ), the same concept is represented with a two-handed sign in which the concept is schematized by tracing the shape of the spoon's curved surface (Figure 10.1b). In Italian Sign Language (LIS), the sign for the same concept is schematized differently, i.e. a one-handed sign is used where the handshape represents the spoon itself combined with the movement towards the mouth (Figure 10.1c). In all three of these examples, the form of the sign resembles the concept of *a*

9780192886491_Book.indb 179 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:40

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

spoon, but they differ regarding the type of semiotic mapping that has been selected for forming the sign. These examples clearly show that iconicity is not an absolute property, but rather it may come in different types where iconicity is conventionalized within each sign language through an analogue-building model.



FIGURE 10.1 Different types of iconic mapping used for SPOON.
a) action-based iconicity in Turkish Sign Language, b) perception-based iconicity in Czech Sign Language, and c) a combination of action- and perception-based iconicity in Italian Sign Language.

This brings us to the notion of *types of iconic representation* as they are essential to determine the semiotic relationship between form and its referent (Padden et al. 2013, 2015; Emmorey 2014; Perniss and Vigliocco 2014; Ortega and Özyürek 2020). Iconic relation between the form of the signs and the meaning can be either *action-based*, i.e. motivated by the motion associated with the concept, or it can be *perception-based*, i.e. motivated by the visual form of the concept (Perniss and Vigliocco 2014). Action-based signs represent how an object moves or how it is held or manipulated (also known as *handling* and *manipulation* strategies) (Figure 10.1a). In turn, perception-based signs either represent the whole object or some salient features of the object using a specific handshape (also known as a *representation* or *object/instrument* strategy) (Figure 10.1c) or by tracing an outline of the object (Figure 10.1b). Furthermore, the form and motion associated with the concept can be also simultaneously combined in a sign (Padden et al. 2013, 2015; Ortega et al. 2014) (Figure 10.1c). Sign languages can have all three types of representation, but they may differ based on which type is used for which concept or semantic category as we see in Figure 10.1. These differences have been considered to indicate typological diversity across sign languages (Padden et al. 2013, 2015).

Padden et al. (2013, 2015) introduce the notion of *patterned iconicity*, suggesting that different types of iconicity may function as a tool in sign languages revealing the semantic class of a group of signs. For example, they show that signers of American Sign Language (ASL) distinguish nouns and verbs related to tool use by using different types of iconicity—perception-based iconicity for nouns and action-based iconicity for verbs (Padden et al. 2015).

Different types of iconicity are particularly relevant for language evolution research as some mappings might be more readily available in communication in the visual modality while others might take time to emerge. For example, research shows that signers of emerging sign languages, such as Nicaraguan Sign Language (NSL), use more action-based signs with their gradual decrease in favour of perception-based signs in later

9780192886491_Book.indb 180 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:41

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

cohorts of language users (Pyers and Senghas 2020). However, iconicity is not only diverse with respect to the different types of iconic mapping but also the degree of the perceptual resemblance between form and its meaning, i.e. some signs can be perceived as more iconic than others. Given that one of the main assumptions in language evolution is the decrease of iconicity in a linguistic system over time, having a method to assess the degree of iconicity of the linguistic forms is particularly useful.

Thus, below, we outline two methods that can be employed to study the evolution of iconicity in sign languages at the lexical level. First, we describe the method used for quantitative assessment of the degree of iconicity known as *iconicity ratings*. Then, we describe a method that employs the categorization of different representation types in established sign languages and silent gesture paradigm (for more on silent gesture methodology, please see Chapter 9, this volume) to provide some insights into the evolutionary trajectory of iconic forms in sign languages.

10.2.2 Iconicity ratings

As the previously described definition of iconicity suggests, what might seem a clearly iconic sign to some might seem only slightly iconic to others, or even not iconic at all. Thus, the extent to which a sign can be considered iconic is somewhat problematic. Furthermore, the perceived iconicity of the signs can be driven by linguistic as well as cultural knowledge. For this reason, the subjective intuitions of individual researchers regarding the presence or absence, or the degree of iconicity of the signs, arguably, is not a reliable measure (Motamedi, Little, et al. 2019; Winter and Perlman 2021).

One crucial method in the study of the iconicity of lexical forms in sign languages in a systematic way is known as *iconicity ratings* which serve as a measure to quantify the degree to which a sign appears to be connected to its meaning (for an in-depth overview on the method of iconicity ratings in language evolution research including spoken languages, see Motamedi, Little, et al. 2019; Winter and Perlman 2021).

Measuring iconicity of single concepts is operationalized in the following way: a large number of people are invited to provide a rating, usually on a Likert scale from 1 ('not iconic at all'/'arbitrary') to 7 ('extremely iconic'/'iconic'), regarding how iconic the specific sign presented as a video recording is (Vinson et al. 2008; Caselli et al. 2017; Caselli and Pyers 2017). In such a way, the degree of iconicity of a sign is rendered more objective considering that the rating can be generalized rather than being limited to the intuition of a single researcher (Vinson et al. 2008; Caselli and Pyers 2017; Motamedi, Little, et al. 2019; Winter et al. 2017; Winter and Perlman 2021). Iconicity ratings thus provide a robust quantitative measure that can be then used to compare the degree of iconicity of linguistic forms used in different sign languages and in different stages of its evolution.

Some research also assesses the *transparency* of the lexical signs by first asking to guess their meaning (Pizzuto and Volterra 2000; Ortega et al. 2019; Sehyr and Emmorey 2019) and/or only then asking for the iconicity rating for the meaning given. Thus, the

9780192886491_Book.indb 181 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:41

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

transparency of a sign reflects whether the meaning of a particular sign can be inferred solely from its lexical form. In turn, the iconicity rating provides a quantitative measure for the strength of the mapping between the lexical sign and its meaning (for a discussion on the difference between iconicity and transparency, see Occhino et al. 2017).

There have also been different approaches regarding the participants from whom the iconicity ratings are collected. Iconicity ratings can be elicited from deaf participants where signers of a specific sign language are asked to rate the iconicity of the signs in the respective language (e.g. Vinson et al. 2008; Trettenbrein et al. 2021). Deaf signers can also be recruited to rate the iconicity of signs of a different sign language (e.g. Adam et al. 2007; Occhino et al. 2017). Finally, hearing participants with no knowledge of any sign language can be used as raters (e.g. Caselli et al. 2017; Sehyr and Emmorey 2019; Trettenbrein et al. 2021). However, comparative studies have shown that the cultural and linguistic experience of the raters mediate perceived iconicity. For example, deaf raters have been shown to rate signs of their language as more iconic than signs of another sign language (Occhino et al. 2017), and hearing raters provide higher iconicity ratings to the signs than signers (Sehyr and Emmorey 2019). Thus, arguably, ratings elicited from hearing raters exclude linguistic experience as a potential bias in the perceived iconicity.

Note that not only linguistic but also cultural knowledge may have an influence on the perceived iconicity of the sign (Pizzuto and Volterra 2000). Let's take an example of the concept of *tea*. The way the tea is served differs vastly across cultures. For example, in Turkish culture, tea is served in a specific glass with a spoon to stir the sugar. In turn, in Japanese culture, tea is served in a cup in which the tea is poured from a specific teapot. In many Western cultures, the most common way to drink tea is using a tea bag dipped in a cup of boiling water. Unsurprisingly, in different sign languages, a sign for TEA is schematized according to the canonical concept of the tea of the specific culture (see Figure 10.2). While all three signs are clearly iconic, the cultural and not the linguistic knowledge might be the determinant in the perceived degree of iconicity of each lexical sign.

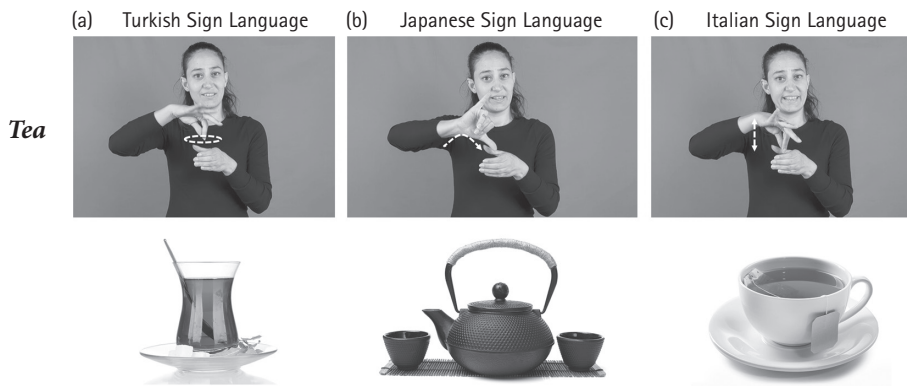


FIGURE 10.2 Lexical signs for TEA in a) Turkish Sign Language, b) Japanese Sign Language, and c) Italian Sign Language.

9780192886491_Book.indb 182 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:41

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Test items for which iconicity ratings are collected are presented in the form of video recordings in which a lexical sign is articulated by a signer. These can be lexical items that have been video recorded specifically to form a lexical database in a specific sign language (e.g. ASL: Caselli et al. 2017; BSL: Vinson et al. 2008; DGS: Trettenbrein et al. 2021). Alternatively, it is also possible to use data from online dictionaries such as Spreadthesigns.com online dictionary (Kimmelman et al. 2018), which currently contains more than 25,000 video-recorded lexical entries per language and spans forty-four sign languages (retrieved on 18 February 2025).

When collecting iconicity ratings of lexical signs from signers of the respective sign language, either an image of the concept is used or there is no image provided as the presented signs form a part of the lexicon of the specific sign language (Sehyr and Emmorey 2019). When collecting iconicity ratings from hearing raters, concepts are represented either as images or as the written translation of the concept (Grote and Linz 2003; Caselli et al. 2017; Caselli and Pyers 2017; Sehyr and Emmorey 2019). Items used for rating elicitation usually span various lexical classes (e.g. nouns, verbs, adjectives) and semantic domains (e.g. manipulable and non-manipulable objects, actions, animals, body parts, colours, etc.), allowing for systematic comparisons (Perlman et al. 2018).

As the concept of iconicity can be confusing, particular attention must be dedicated to how participants are instructed to rate the concepts. Ratings can be elicited by asking participants to rate *to what extent the signs look like what they mean* (e.g. Vinson et al. 2008; Sehyr and Emmorey 2019), or *how much each sign looks like its translation* when presented with the written translation for the sign (e.g. Caselli et al. 2017). Furthermore, before testing, iconicity can be explained better by providing examples of different signs that vary in the degree of iconicity as well as presenting clearly arbitrary signs (Vinson et al. 2008; Sehyr and Emmorey 2019). Iconicity ratings can be collected by inviting participants to provide ratings in person in the lab, but they can also be collected online (Motamedi, Little, et al. 2019).

So far, iconicity ratings for large sets of vocabulary items have been collected only in a few sign languages: American Sign Language (Lieberth and Gamble 1991; Caselli et al. 2017; Occhino et al. 2017; Sehyr and Emmorey 2019); British Sign Language (Vinson et al. 2008; Thompson et al. 2012); and German Sign Language (Grote and Linz 2003; Occhino et al. 2017; Trettenbrein et al. 2021). Such databases make it possible to assess the distribution of iconicity in the vocabularies of these sign languages as well as systematic comparisons between them and even between sign and spoken languages (Perlman et al. 2018; Thompson et al. 2020).

While iconicity ratings have been mostly used in language acquisition and processing research, operationalizing this method in the language evolution field constitutes a promising endeavour for future research. For example, Caselli and Pyers (2020), who focused on the role of iconicity types (i.e. *patterned iconicity*) and degree of iconicity (i.e. iconicity ratings) in language acquisition, have shown that action-based signs (termed *pantomimic* in the study) are rated higher for their iconicity than other types of signs. This finding lends empirical support to previous research suggesting that

9780192886491_Book.indb 183 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:41

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

action-based signs, being more imitative, should provide a more direct iconic mapping than more schematized and abstracted perception-based signs (Emmorey 2014; Ortega et al. 2014). Accordingly, investigating the iconic representation types concerning the degree of their perceived iconicity across different population groups might prove to be useful in better understanding the relationship between iconicity, schematization, and abstraction in language evolution.

10.2.3 Evolution of iconicity: Sign languages vs. silent gesture

A different way to study patterned iconicity in sign language evolution research is to first establish a baseline for the iconicity. An experimental paradigm, known as the *silent gesture* (Goldin-Meadow et al. 2008), serves as such a baseline given that it is representative of the iconic potential for communication without a linguistic system and, as such, reflects the general affordances of semiosis in the visual modality. The silent gesture paradigm has been used for eliciting iconic representations of single concepts that can then be compared to how the same concepts are represented in established sign languages (Schembri et al. 2005; Padden et al. 2013, 2015; Brentari et al. 2015; Hwang et al. 2017; Ortega and Özyürek 2020). For example, research comparing signers and silent gesturers regarding iconicity types used to represent single concepts shows that while silent gesturers prefer using action-based iconicity (handling/acting) to represent hand-held tools (Padden et al. 2013, 2015), signers also use a combination of action- and perception-based iconicity type by also representing the tool itself (Padden et al. 2013, 2015). This finding suggests that some iconicity types might be more salient for expressing meaning in the visual modality while other types, for example ones that require more abstraction and schematization as is the case with perception-based signs, might be systematically recruited later in language evolution (Emmorey 2014).

Systematic comparison concerning iconicity types used for describing the same concept between signers and silent gesturers can reveal not only differences but also common cognitive biases regarding how visual modality affords representation of meaning (Schembri et al. 2005; Padden et al. 2013, 2015; Brentari et al. 2015; Hwang et al. 2017; Ortega and Özyürek 2020). For example, Padden et al. (2015) show that similar to signers, silent gesturers also differentiate between action-based and perception-based strategies to differentiate between objects and actions. Gesturers, however, lack the systematicity that can be observed in signers. To elicit nouns, authors used images depicting objects while to elicit verbs they used videos showing how the respective object is used by a person.

Similar findings have been also found in agentive vs. non-agentive events (i.e. a stimulus depicting an item moved by an agent vs. an item moving on its own), where agentive events elicited more action-based iconic representations in both signers and gesturers, while non-agentive events elicited more perception-based iconicity in both

9780192886491_Book.indb 184 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:41

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

groups (Brentari et al. 2015). These findings point towards gestural origins in selecting specific semiotic mappings for communicating about specific semantic and lexical categories which systematize in language evolution.

Thus, it becomes evident that to understand sign language evolution, it is important to tease apart which forms and structures that we see in sign languages are driven by general affordances in communication in the visual modality, and which forms and structures are a result of evolutionary change of language. A comparison of iconic structures in silent gestures and established sign languages constitutes a method to achieve this.

10.3 Method to study the evolution of iconicity beyond the lexical level

10.3.1 Simultaneity and diagrammatic iconicity

When it comes to iconicity beyond the lexical level, it is important to differentiate between two types: 1) *imagistic iconicity*, which is based on the perceptual resemblance between the meaning elements and the linguistic form (as described above for the lexical level); and 2) *diagrammatic iconicity*, which is based on the structural resemblance of the relationship between the meaning elements and the relationship between linguistic forms used to represent these elements (Perniss 2007; Slonimska et al. 2020). Perceptually, each element encoded by different articulators can be imagistically iconic regarding the meaning it is referring to. For example, in Figure 10.3, a signer is using an iconic and simultaneous construction to represent a cat gazing at, holding, and petting a little bear in its hand. The mapping between the torso, head, and eyes of the signer imagistically represents the cat and its gaze, and the signs used for representing holding and petting are imagistically iconic to the actions of holding and petting. However, the spatial organization of the body articulators (and the respective meanings they are encoding) is representative of diagrammatic iconicity as the way one articulator is positioned with respect to the others maps onto the spatial and/or temporal relationship between meaning elements. Namely, the encoded spatial and temporal relation of the articulators maps onto the spatial and temporal relation between meaning elements of the event. The use of iconicity for both, imagistic and diagrammatic mapping, allows a more direct and thus more iconic representation of an event through simultaneity, considering that perceptually simultaneous elements of an event can be represented as such in a linguistic system (Slonimska et al. 2022).

The mainstream view in language evolution, including sign languages, is that languages, as they evolve, shift from iconic and simultaneous forms and structures to arbitrary and sequential ones (Senghas et al. 2004; Aronoff et al. 2005). However, this has not been tested at discourse levels and with events containing interaction

9780192886491_Book.indb 185 /12_first_proofs/first_pr…/xml_for_typesetting 27-Mar-25 23:02:41

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

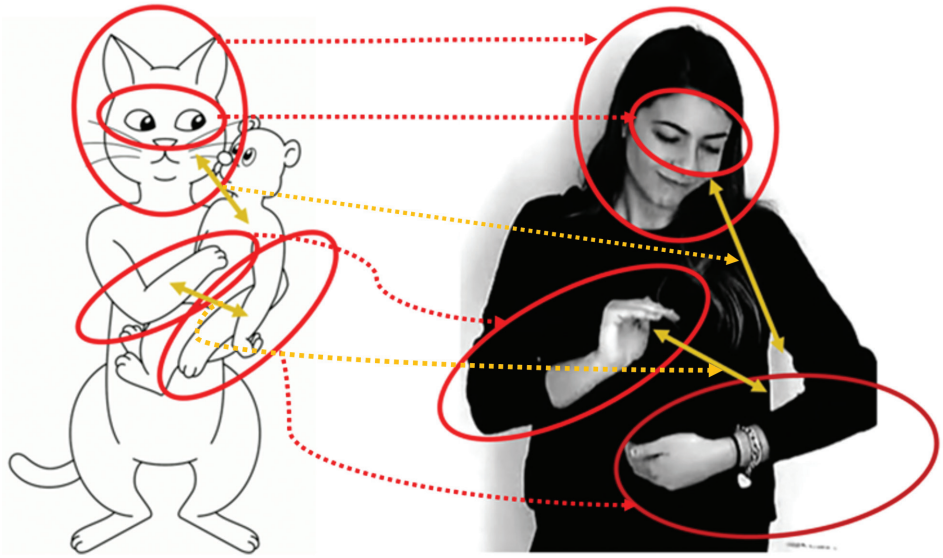


FIGURE 10.3 An iconic and simultaneous construction in Italian Sign Language of a cat gazing at, holding, and petting a bear.

Individual meaning elements are mapped through *imagistic* iconicity onto different body articulators, marked by red circles. The temporal and spatial relationship between meaning elements is mapped through *diagrammatic* iconicity by positioning body articulators in space in respective relationship to each other.

between multiple agents—where simultaneous structures using multiple articulators and diagrammatic iconicity are more likely to be recruited. In a recent study, we tested whether diagrammatic iconicity that can be expressed through the use of simultaneous articulators is a feature that evolves in a language or is there from the beginning as an affordance of visual modality.

10.3.2 Evolution of simultaneity: Sign languages vs. silent gesture

To understand the role of diagrammatic iconicity in language evolution, we compared silent gesture with sign language structures. We hypothesized that if simultaneous and iconic structures have evolved to maximize expressive capacity in sign languages, signers (in our study, we focused on Italian Sign Language) should use such constructions more than silent gesturers when having to describe complex events involving multiple referents and actions (Slonimska et al. 2022). We found that even though visual modality alone allows iconic and simultaneous structures, they are recruited much less in silent gesture compared to sign language—suggesting their emergence as a linguistic property later—rather than being a strategy arising in the initial stages of the language evolution process. Below, we describe the methodology used in the study.

9780192886491_Book.indb 186 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:41

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

The design of the experiment consisted of the systematic manipulation of the information density of an event, quantified as the specific number of semantic information units requiring encoding. The stimuli presented were black-and-white cartoon images. Events represented in the images involved two animate referents and were manipulated in such a way that the information density of these events was systematically increased, requiring participants, whose task was to describe these images to another person (a confederate), to be more communicatively efficient as the information density of the events increased. There were five information density levels in total (see Figure 10.4) and there were six different referent pairs yielding thirty test items in total.

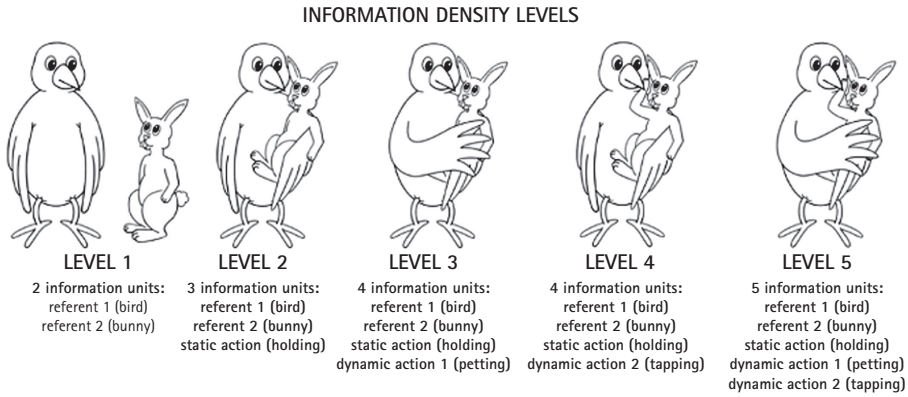


FIGURE 10.4 Example of five *information density levels* for the referent pair: 'bird' and 'bunny'.
The format of the images in levels 1 and 2 is PNG and the format of the images in levels 3, 4, and 5 is GIF. In the GIFs, the dynamic action of the referents is animated. Original (animated) stimuli are available at https://osf.io/g57p2.

Such an experimental design allowed us to assess the effect of information density of the event on simultaneous and iconic constructions as well as the iconic strategies used by signers and silent gesturers in a systematic way given that the number of information units requiring encoding was fixed and controlled.

Each participant was informed that they would play a *director–matcher* game and that they had been assigned the role of *director*. Deaf participants were instructed in LIS by a native LIS signer and hearing participants were instructed in Italian by a native or highly proficient Italian speaker. Furthermore, hearing participants were instructed that to describe the images they could only use their hands and body without any speech. Participants were informed that the director's task was to describe the images appearing on the TV screen to the matcher so the matcher, who was sitting in front of the participant and could not see the TV screen could choose the correct image on their laptop. All images were displayed one by one in a semi-randomized order so that the same information density level and the same referent pair never appeared one after another.

The participants' productions were video recorded and used for coding in ELAN. LIS data were coded for the number of movement segments (MS), the number of encoded information units in each movement segment, as well as the linguistic strategy or strategies used in each movement segment. Silent gesture data were

9780192886491_Book.indb 187 /12_first_proofs/first_pr[illegible]/xml_for_typesetting 27-Mar-25 23:02:42

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

coded for movement segments (MS), the number of encoded information units in each movement segment, and the type of information unit encoded by each articulator. These measures were analysed using generalized mixed-effects models to test whether both groups were different regarding the amount of iconic and simultaneous constructions used.

Results revealed that iconic and simultaneous constructions used by signers were used more frequently and they were informationally denser than those used by silent gesturers. Qualitative analyses revealed that while signers employed diagrammatic iconicity to combine multiple related meaning elements in simultaneous structures, silent gesturers were more limited in their encodings due to relying primarily on imagistic iconicity (Figure 10.5). The findings of this study suggest that highly iconic and simultaneous constructions used by signers (Figure 10.5b) constitute a linguistic property that has evolved to maximize expressive capacity in sign languages.

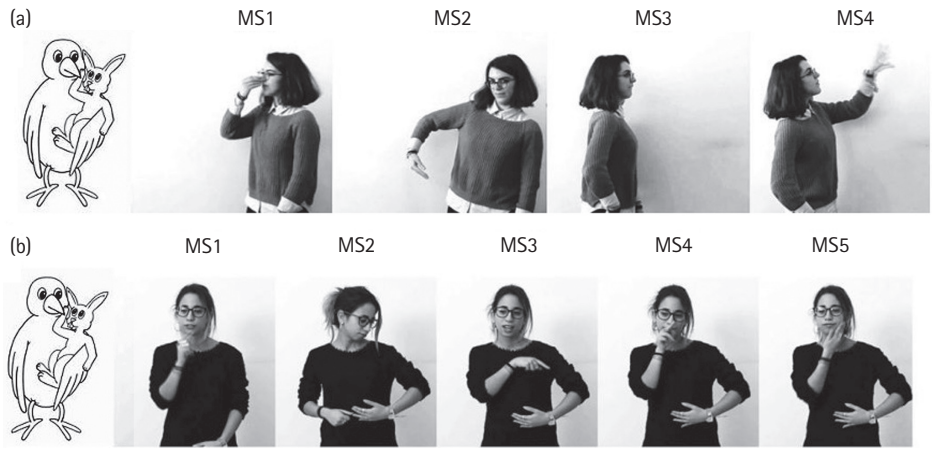


FIGURE 10.5 Descriptions of an image (*information density level 4*) produced by a) a hearing participant using silent gesture and b) a signer using Italian Sign Language.

An interesting follow-up question is whether such simultaneous structures are perceived as more iconic than segmented ones. Iconicity ratings for concepts beyond the lexical level so far have not been studied. More generally, the role of diagrammatic iconicity and simultaneity in the sign language field as well as language evolution has been less often studied (but see Vermeerbergen et al. 2007). However, once a hypothesis of evolving iconic and simultaneous structures in visual linguistic systems is introduced, it raises new questions to be explored, such as which iconic and simultaneous structures are available from the initial stages of language emergence, which structures evolve, why, and how. While methodology to tackle these questions is still in its infancy, it has the potential to provide a more nuanced understanding of the evolutionary trajectory of sign languages by accounting for the evolution of the visual modality-specific structures and thus challenging the longstanding notion in language evolution of the gradual drift towards arbitrariness and linearization.

9780192886491_Book.indb 188 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:42

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

10.4 Future directions in iconicity and sign language evolution research: Computer vision techniques

Another promising method for studying the evolution of iconicity in sign languages is recently developing computer vision techniques that allow quantifying the kinematic properties of signs and gestures. So far, only one study has applied this method in language evolution research. Namely, Pouw, Dingemanse, et al. (2021), who analysed the kinematics of a dataset based on a study using an iterated learning paradigm in which silent gestures produced by one group of participants were then used as a training set for the next group of participants, leading to a chain of simulated 'generations of language learners' (Motamedi, Schouwstra, et al. 2019). The authors used OpenPose (Cao et al. 2017), which is a pre-trained deep neural network used for analyses of the human body based on the video data to assess the following measures: sub-movements, intermittency, gesture space, rhythm, and temporal variability (or rhythmicity). These measures were used to assess the kinematic quality of the gestures. Authors found that, as the silent gestural systems evolve, the gestures become smaller, smoother, and more rhythmic.

Furthermore, Pouw, de Wit, et al. (2021) have devised a method that allows assessing whether semantically similar iconic gestures are also similar regarding their kinematic properties. This is achieved by combining semantic word2vec distances with kinematic distances of the gestures using a Dynamic Time Warping (DTW) technique. The authors show the existence of a dependency between kinematic features and the semantic space of the concepts. While these methods are yet to be applied to studying the evolution of iconicity in sign languages, targeting kinematic features of iconic signs can be a promising direction of research in the future.

10.5 Conclusion

In this chapter, we described some methods that have been, and may be employed for understanding the evolution of iconicity at the lexical level and its use through simultaneity at the discourse level in sign languages. It is a matter of debate in sign language evolution whether iconicity is a property that helps start up a linguistic system from gesture, which then erodes over time, or if structures using iconicity are those that evolve. We highlighted that the comparison of iconic and simultaneous structures in established sign languages and structures in silent gesture serves to tease apart which structures in established sign languages are available as general affordances of communicating meaning in the visual modality and which structures have emerged

9780192886491_Book.indb 189 /12_first_proofs/first_pr…/xml_for_typesetting 27-Mar-25 23:02:42

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

as language evolved. We also described recent methodologies for studying iconicity at the lexical and discourse levels. On this, higher-level iconicity is directly linked to the affordance of representing events more iconically through capturing perceptual simultaneity of event elements. We also outlined newly emerging methods of kinematic analyses that are promising in providing more insights into the evolution of sign languages in a more automated and fine-grained manner. Currently, we have some evidence for whether, and how, iconicity evolves in sign languages when different types of iconicity at different levels of linguistic structure are considered. The methodology offered in this chapter provides various avenues for further research on this topic.

Acknowledgements

We would like to thank Berna Dikyuva for creating Figures 10.1 and 10.2, and the participants of the studies in Slonimska et al. (2020, 2022) for their permission to use their recordings for demonstration purposes (Figures 10.3 and 10.5).

References

Adam, M., Iversen, W., Wilkinson, E., and Morford, I. P. (2007). 'Meaning on the one hand', *Insistent Images* 5: 211–227.

Aronoff, M., Meir, I., and Sandler, W. (2005). 'The paradox of sign language morphology', *Language* 81(2): 301–344.

Arbib, M. A. (2005). 'From monkey-like action recognition to human language: An evolutionary framework for neurolinguistics', *Behavioral and Brain Sciences* 28(2): 105–124.

Brentari, D., Di Renzo, A., Keane, J., and Volterra, V. (2015). 'Cognitive, cultural, and linguistic sources of a handshape distinction expressing agentivity', *Topics in Cognitive Science* 7(1): 95–123, https://doi.org/10.1111/tops.12123.

Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. (2017). 'Realtime multi-person 2D pose estimation using part affinity fields', *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 1302–1310, https://doi.org/10.1109/CVPR.2017.143.

Caselli, N. K. and Pyers, J. E. (2017). 'The road to language learning is not entirely iconic: Iconicity, neighborhood density, and frequency facilitate the acquisition of sign language', *Psychological Science* 28(7): 979–987.

Caselli, N. K. and Pyers, J. E. (2020). 'Degree and not the type of iconicity affects sign language vocabulary acquisition', *Journal of Experimental Psychology: Learning, Memory, and Cognition* 46(1): 127.

Caselli, N. K., Sehyr, Z. S., Cohen-Goldberg, A. M., and Emmorey, K. (2017). 'ASL-LEX: A lexical database of American Sign Language', *Behavior Research Methods* 49: 784–801.

Cuxac, C. and Sallandre, M. A. (2007). 'Iconicity and arbitrariness in French Sign Language: Highly iconic structures, degenerated iconicity and diagrammatic iconicity', *Empirical Approaches to Language Typology* 36: 13.

Dingemanse, M. (2011). 'Ideophones and the aesthetics of everyday language in a West-African society', *Senses and Society* 6(1): 77–85, https://doi.org/10.2752/174589311X12893982233830.

9780192886491_Book.indb 190 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:42

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Dingemanse, M. (2015). 'Ideophones and reduplication: Depiction, description, and the interpretation of repeated talk in discourse', *Studies in Language. International Journal Sponsored by the Foundation 'Foundations of Language'* 39(4): 946–970, https://doi.org/10.1075/sl.39.4.05din.

Emmorey, K. (2014). 'Iconicity as structure mapping', *Philosophical Transactions of the Royal Society B: Biological Sciences* 369(1651): 20130301, https://doi.org/10.1098/rstb.2013.0301.

Goldin-Meadow, S., So, W. C., Özyürek, A., and Mylander, C. (2008). 'The natural order of events: How speakers of different languages represent events nonverbally', *Proceedings of the National Academy of Sciences of the United States of America* 105(27): 9163–9168, https://doi.org/10.1073/pnas.0710060105.

Grote, E. and Linz, K. (2003). 'The influence of sign language iconicity on semantic conceptualization', in W. G. Müller and O. Fischer (eds.), *From Sign to Signing* (Vol. 3), 23–40. John Benjamins Publishing.

Hwang, S.-O., Tomita, N., Morgan, H., et al. (2017). 'Of the body and the hands: Patterned iconicity for semantic categories', *Language and Cognition* 9(4): 573–602.

Jackendoff, R. (2010). 'Your theory of language evolution depends on your theory of language', in R. K. Larson, V. Déprez, and H. Yamakido (eds.), *The Evolution of Human Language: Biolinguistic Perspectives. Approaches to the Evolution of Language*, 63–72. Cambridge: Cambridge University Press.

Kimmelman, V., Klezovich, A., and Moroz, G. (2018). 'IPSL: A database of iconicity patterns in sign languages. Creation and use', *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Kyle, J. G., Kyle, J., and Woll, B. (1988). *Sign Language: The Study of Deaf People and Their Language*. Cambridge: Cambridge University Press.

Lieberth, A. K. and Gamble, M. E. B. (1991). 'The role of iconicity in sign language learning by hearing adults', *Journal of Communication Disorders* 24(2): 89–99.

Lutzenberger, H., Mudd, K., Stamp, R., et al. (2023). 'The social structure of signing communities and lexical variation: A cross-linguistic comparison of three unrelated sign languages', *Glossa: A Journal of General Linguistics* 8(1).

Motamedi, Y., Little, H., Nielsen, A., and Sulik, J. (2019). 'The iconicity toolbox: Empirical approaches to measuring iconicity' *Language and Cognition* 11(2): 188–207.

Motamedi, Y., Schouwstra, M., Smith, K., et al. (2019). 'Evolving artificial sign languages in the lab: From improvised gesture to systematic sign', *Cognition* 192: 103964, https://doi.org/10.1016/j.cognition.2019.05.001.

Occhino, C., Anible, B., Wilkinson, E., and Morford, J. P. (2017). 'Iconicity is in the eye of the beholder: How language experience affects perceived iconicity', *Gesture* 16(1): 100–126, https://doi.org/10.1075/gest.16.1.040cc.

Ortega, G. and Özyürek, A. (2020). 'Types of iconicity and combinatorial strategies distinguish semantic categories in silent gesture across cultures', *Language and Cognition* 12(1): 84–113, https://doi.org/10.1017/langcog.2019.28.

Ortega, G., Sumer, B., and Özyürek, A. (2014). 'Type of iconicity matters: Bias for action-based signs in sign language acquisition', *36th Annual Meeting of the Cognitive Science Society (CogSci 2014)* 1114–1119.

Ortega, G., Schiefner, A., and Özyürek, A. (2019). 'Hearing non-signers use their gestures to predict iconic form-meaning mappings at first exposure to sign', *Cognition* 191: 103996, doi:10.1016/j.cognition.2019.06.008.

Padden, C. A., Meir, I., Hwang, S.-O., et al. (2013). 'Patterned iconicity in sign language lexicons', *Gesture* 13(3): 287–308, https://doi.org/10.1075/gest.13.3.03pad.

9780192886491_Book.indb 191 /12_first_proofs/first_pr/xml_for_typesetting 27-Mar-25 23:02:42

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Padden, C., Hwang, S.-O., Lepic, R., and Seegers, S. (2015). 'Tools for language: Patterned iconicity in sign language nouns and verbs', *Topics in Cognitive Science* 7(1): 81–94, https://doi.org/10.1111/tops.12121.

Perlman, M., Little, H., Thompson, B., and Thompson, R. L. (2018). 'Iconicity in signed and spoken vocabulary: A comparison between American Sign Language, British Sign Language, English, and Spanish', *Frontiers in Psychology* 9: 1433.

Perniss, P. M. (2007). *Space and Iconicity in German Sign Language (DGS)* [[S.l.]: MPI Series in Psycholinguistics], https://repository.ubn.ru.nl/handle/2066/30937.

Perniss, P. and Vigliocco, G. (2014). 'The bridge of iconicity: From a world of experience to the experience of language', *Philosophical Transactions of the Royal Society B: Biological Sciences* 369(1651): 20130300, https://doi.org/10.1098/rstb.2013.0300.

Pizzuto, E., and Volterra, V. (2000). 'Iconicity and transparency in sign languages: A cross-linguistic cross-cultural view', in K. Emmorey and H. Lane (eds.), *The Signs of Language Revisited: An Anthology to Honor Ursula Bellugi and Edward Klima,* 261–286. Mahwah, NJ, London: Lawrence Erlbaum Associates Publishers.

Pouw, W., de Wit, J., Bögels, S., et al. (2021). 'Semantically related gestures move alike: Towards a distributional semantics of gesture kinematics', *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. Human Body, Motion and Behavior: 12th International Conference, DHM 2021, Held as Part of the 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings, Part I*, 269–287.

Pouw, W., Dingemanse, M., Motamedi, Y., and Özyürek, A. (2021). 'A systematic investigation of gesture kinematics in evolving manual languages in the lab', *Cognitive Science* 45(7): e13014.

Pyers, J. and Senghas, A. (2020). 'Lexical iconicity is differentially favored under transmission in a new sign language: The effect of type of iconicity', *Sign Language and Linguistics* 23(1–2): 73–95, https://doi.org/10.1075/sll.00044.pye.

Sandler, W. (2018). 'The body as evidence for the nature of language', *Frontiers in Psychology* 9, https://www.frontiersin.org/articles/10.3389/fpsyg.2018.01782.

Schembri, A., Jones, C., and Burnham, D. (2005). 'Comparing action gestures and classifier verbs of motion: Evidence from Australian Sign Language, Taiwan Sign Language, and nonsigners' gestures without speech', *Journal of Deaf Studies and Deaf Education* 10(3): 272–290.

Sehyr, Z. S. and Emmorey, K. (2019). 'The perceived mapping between form and meaning in American Sign Language depends on linguistic knowledge and task: Evidence from iconicity and transparency judgments', *Language and Cognition* 11(2): 208–234.

Senghas, A., Kita, S., and Özyürek, A. (2004). 'Children creating core properties of language: Evidence from an emerging sign language in Nicaragua', *Science* 305(5691): 1779–1782, https://doi.org/10.1126/science.1100199.

Sidhu, D. M., Vigliocco, G., and Pexman, P. M. (2020). 'Effects of iconicity in lexical decision', *Language and Cognition: An Interdisciplinary Journal of Language and Cognitive Science* 12(1): 164–181, https://doi.org/10.1017/langcog.2019.36.

Slonimska, A., Özyürek, A., and Capirci, O. (2020). 'The role of iconicity and simultaneity for efficient communication: The case of Italian Sign Language (LIS)', *Cognition* 200: 104246, https://doi.org/10.1016/j.cognition.2020.104246.

Slonimska, A., Özyürek, A., and Capirci, O. (2022). 'Simultaneity as an emergent property of efficient communication in language: A comparison of silent gesture and sign language', *Cognitive Science* 46(5): e13133.

9780192886491_Book.indb 192 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:42

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Taub, S. F. (2001). *Language from the Body: Iconicity and Metaphor in American Sign Language*. Cambridge: Cambridge University Press, https://doi.org/10.1017/CBO9780511509629.

Thompson, B., Perlman, M., Lupyan, G., et al. (2020). 'A data-driven approach to the semantics of iconicity in American Sign Language and English', *Language and Cognition* 12(1): 182–202.

Thompson, R. L., Vinson, D. P., Woll, B., and Vigliocco, G. (2012). 'The road to language learning is iconic: Evidence from British Sign Language', *Psychological Science* 23(12): 1443–1448, https://doi.org/10.1177/0956797612459763.

Tomasello, M. (2008). *Origins of Human Communication*. Cambridge, MA: MIT Press.

Trettenbrein, P. C., Pendzich, N.-K., Cramer, J.-M., et al. (2021). 'Psycholinguistic norms for more than 300 lexical signs in German Sign Language (DGS)', *Behavior Research Methods* 53: 1817–1832.

Vermeerbergen, M., Leeson, L., and Crasborn, O. A. (2007). *Simultaneity in Signed Languages: Form and Function* (Vol. 281). Amsterdam: John Benjamins Publishing.

Vinson, D. P., Cormier, K., Denmark, T., et al. (2008). 'The British Sign Language (BSL) norms for age of acquisition, familiarity, and iconicity', *Behavior Research Methods* 40: 1079–1087.

Vinson, D., Thompson, R. L., Skinner, R., and Vigliocco, G. (2015). 'A faster path between meaning and form? Iconicity facilitates sign recognition and production in British Sign Language', *Journal of Memory and Language* 82: 56–85, https://doi.org/10.1016/j.jml.2015.03.002.

Winter, B. and Perlman, M. (2021). 'Iconicity ratings really do measure iconicity, and they open a new window onto the nature of language', *Linguistics Vanguard* 7(1): 20200135, https://doi.org/10.1515/lingvan-2020-0135.

Winter, B., Perlman, M., Perry, L. K., and Lupyan, G. (2017). 'Which words are most iconic? Iconicity in English sensory words', *Interaction Studies* 18(3): 443–464.

9780192886491_Book.indb 193 /12_first_proofs/first_pr…/xml_for_typesetting 27-Mar-25 23:02:42

CHAPTER 11

THE STUDY OF CREOLE LANGUAGES

DAMIAN BLASI, FELICIA BISNATH, AND PUI YIU SZETO

11.1 The notion of Creole languages and its social, historical, and political underpinnings

The term 'Creole' comes from the Portuguese and Spanish word meaning 'raised in the home'. Within the Caribbean, the term was initially used to distinguish humans, animals, plants, traditions, etc., originating in European colonies from those of the colonizing country (DeGraff 2005; Velupillai 2015, and references within). The term 'Creole' has varied uses depending on the field. In linguistics, a 'Creole' is a class of natural languages which arose through intense contact and is used as a first language by an entire community. Typically, the label refers to languages that developed during the transatlantic slave trade, e.g. Haitian Creole; however, many Creoles with other contributors are known, e.g. Sri Lanka Malay. In fields viewed as more peripheral to Creole linguistics, like anthropology, derivations of the term like 'creolization' are used to refer to general language mixture and variation (Eriksen 2007; Palmié 2016). Outside of academia, Creole users vary in how they refer to their languages. For example, Cabo Verdean Creole and Dominican French Creole are known as Kriolu and Kwéyòl, respectively, which are near equivalents of English 'Creole'. Alternatively, users of Trinidadian English Creole and Trinidadian French Creole use terms like 'dialect' and 'patwa' respectively (Bancu et al. 2024). The use of terms like 'dialect' point to the historical minoritization of Creoles in the communities they are used in and to their related minoritization in linguistics (Rickford 1985; DeGraff 2001, 2005, 2020; Bouchard 2022). Note, however, that within and across Creole-using communities, attitudes vary (Wassink 1999; Romaine and Rickford 1999; Nero 2014; Mühleisen 2022).

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

In practice, Creoles are distinguished from other contact languages by their social contexts of emergence. A typical Creole ecology is that of the plantations associated with the transatlantic slave trade, in which a relatively small population of European colonizers and their language(s) dominated a heterogeneous and larger multilingual population (Ansaldo and Szeto 2020). The language of the colonizers is referred to as the 'lexifier' or 'superstrate', in recognition of its significant impact on the form of the lexicon of the emerging Creole language. The subjugated group can be very heterogeneous, especially when they are transplanted to the territory. Their varieties are usually referred to as 'substrate' languages. In classic Creole contexts, the lexifier and the substrates differ significantly in their grammar. For instance, in Mauritian Creole, the major lexifier is French and other contributors include: Malagasy, Tamil, Wolof, Mandinka, Bantu languages of East Africa, Bhojpuri, and English (Baker and Kriegel 2013). Other groups holding varying levels of power may also be present in the classic context, and may contribute to further linguistic diversity, sometimes in the form of mixed codes (Lim and Ansaldo 2016).

Apart from the plantation setup, other demographic and ecological setups have been identified in the process of Creole genesis (Arends 1995; Velupillai 2015). These include the fort ecology, where large and heterogeneous populations of enslaved people are kept for certain periods of time and the homestead ecology, where small groups of masters and servants live in close proximity to each other. The plantation ecology exemplifies an exogenous context in which *all* participating groups are immigrants e.g. Gullah Geechee (southern USA), Fa d'Ambô (Gulf of Guinea). Contrastingly, the fort ecology exemplifies an endogenous context in which it is *not necessarily* the case that the groups involved are immigrants, e.g. Kinubi (Kenya, Uganda). A crucial distinction between these two is that in the former, there is less contact with substrate languages while in the latter there is protracted contact.

Beyond these well-studied cases, other languages have been framed, sometimes controversially, as Creoles. This is in part due to differences in how Creole is defined and covert and overt ideological trends, e.g. strategic analogy between Creoles and sign languages to argue that the latter are true languages (Lupton and Salmons 1996). For instance, proposals have been made for and against various varieties of English, e.g. Middle English, African American English, and Aboriginal English being Creoles or Creole-derived (Danchev 1997 and references within; McWhorter 2002; Pennycook 2007; Rickford 2021). American Sign Language (ASL) has also been argued to be a Creole (Woodward 1978; Fischer 1978; Kegl 2008; cf. Lupton and Salmons 1996). Furthermore, Sanogo and Kamei (2019) claim that Langue des Signes d'Afrique Francophone (LSAF), found in Francophone countries in West and Central Africa, is a Creole. Creolization framing has also been applied to entire groups of languages such as the Sinitic languages, which have emerged under very different contexts of intense contact than the Atlantic Creoles (DeLancey 2014).

Most well-known Creole languages (e.g. the English-lexified Pichi, French-lexified Haitian Creole, Portuguese-lexified Cape Verdean Creole, and Spanish-lexified Papiamentu) are found in the Atlantic region due to the transatlantic slave

9780192886491_Book.indb 196 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:43

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

trade. There has been a tendency to overemphasize Creoles from this particular region, thereby unduly generalizing the salient social and linguistic characteristics of these languages (Michaelis 2020). However, Creoles are found in all continents and with a considerable range of histories—from the Pacific, Indian Ocean, to the Arab world, southern Africa, Australia, and beyond (Ansaldo and Meyerhoff 2020). For instance, non-European-lexified Creoles in South and Southeast Asia were not formed in 'prototypical' ecologies involving slave trades or associated with social conditions characterized by subjugation, suppression, or disadvantage (Lee 2020). A concrete example is Nagamese, an Assamese-lexified Creole spoken in the Assam province of India, which is characterized by SOV order and postpositions, both of which do not conform to the linguistic traits assumed to be shared by all Creoles (see sections 11.2 and 11.3).

In spite of these definitional and ecological differences, there is consensus among linguists that Creoles, like any other natural language, are fully-fledged languages that meet the linguistic needs of the communities they are used in. In contrast, the notion of 'pidgin' is often presented as a counterpart to that of Creoles. Pidgins also emerge in intense contact situations (sometimes coincident to those of Creoles), but unlike Creoles, are regarded to be limited in terms of their expressive capacity. They are often discussed along with Creoles for historical reasons, as in the next section.

11.2 The bioprogram hypothesis and related ideas

By far, within the field of language evolution, the most popular idea about the emergence of Creole languages is that of the 'bioprogram hypothesis' (Bickerton 1984). This was first put forward by D. Bickerton in the 1980s, heavily grounded on Hawaiian Pidgin and Hawaiian Creole. These systems flourished in the nineteenth century due to the exploitation of sugar cane plantations in the island, which often saw indentured workers from Asia and the Pacific, the Americas, West Africa, and Europe sharing not only their daily work but also their living premises.

The starting point of this hypothesis is the alleged scarcity of input in typical children growing up in the context of Creole genesis which is insufficient to guarantee the typical process of language transmission characterizing other, non-Creole, languages. Even more, the main communication system overheard by children is (assumedly) not a fully-fledged natural language but a modest pidgin. Pidgins are typically framed as limited codes that serve the most immediate needs of the multilingual society where it is used, and which do not possess any native users. Importantly, the use of the pidgin was subject to substantial variation that seemed to follow the first languages of the individuals using them. For instance, Filipino users of Hawaiian Pidgin will often deploy a VSO order (found in Tagalog and other languages of the island), whereas Japanese users will use SOV instead (typical of Japonic languages). However, within a few generations, a new

9780192886491_Book.indb 197 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:43

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Creole language emerged in the community, leading to the claim that the main basis of the Creole was the preceding pidgin.

This new Creole language contrasts with the pidgin in two ways: (1) no more substantive variation in grammar as before; and (2) striking similarities with other Creole languages around the world. According to Bickerton's analysis, the mismatch between the complexity and regularity of the Creole grammar and the limited functionality of the pidgin could only be explained by appealing to a strong cognitive prior on the nature of language. This prior, presumed to be available to all humans during early childhood, connects the dots between scarce and limited linguistic behaviour, on the one hand, and a full mental model of a language, on the other. Importantly, this prior is not simply biased towards any natural language, but privileges a certain type of grammar above others. In particular, Bickerton preliminarily suggested that this 'default' grammar will be that of a SVO language with little inflectional morphology, among other properties.

While, in creolistics, the 'bioprogram hypothesis' has long gone out of fashion, a number of models preserve some of its main features (e.g. McWhorter 2001; Parkvall 2008; Bakker et al. 2011; Goldshtein 2017). These agree with Bickerton in that the input available to children is limited and that it comes mostly in the form of a pidgin, which then becomes the basis of the Creole. However, when it comes to explaining the recurrence of certain structures in Creoles, they appeal not to a biologically endowed prior on grammars, but on the need for communication. Creoles are 'efficient' languages shaped by the need for communication, before they start gaining the typical opacity and strategies viewed as cumbersome that characterize languages that have been used for a much more extended period of time (e.g. the opacity of certain verbal and nominal forms in English).

11.3 The emergence of Creole languages

No matter how appealing these and other theoretical models of the emergence of Creoles could be, one major limitation in the field is the relatively modest amount of data on the very process of Creole genesis. Because of this, the best available evidence is (1) the description of the Creole languages themselves, and (2) some, sometimes coarse, information on the lexifiers and substrates involved in their development. This has been the foundation of many studies on the emergence of Creole languages based on cross-linguistic comparisons. However, these studies—which oftentimes agree on their conclusions—differ in the languages and the analyses.

This motivated a large-scale assessment of Creole languages by Blasi et al. (2017). Importantly, in order to avoid any bias in the selection of features, they analysed all the data (both in terms of languages and linguistic features) that were available to them. The authors analysed forty-eight diverse Creole languages on over forty-eight

9780192886491_Book.indb 198 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:43

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

different linguistic dimensions (ranging from clausal syntax, phonology, vocabulary, the presence of linguistic categories, and others) and compared them against their putative lexifiers and substrates. First, they found that most Creoles did seem to conform to a distinguishable type of grammar with some salient features, including the identical coding of instrumental and comitative arguments (as in English, 'I walk **with** my friend' = 'I'll fix this **with** a hammer'), and the presence of serial verb constructions. However, Creoles that do not conform to this pattern are those which were the least 'typical' in that they did not involve a combination of European lexifier plus a majority of West African peoples contributing substrate languages. Moreover, the features that most efficiently distinguished Creole from non-Creole languages were not obviously associated with more 'efficient' communication but, instead, were particularly prominent in Europe and West Africa. In particular, the reason Creoles are often regarded as having little morphological inflection is due to the dominant presence of West African languages (which also lack morphological inflection), and further reinforced by the fact that the European lexifiers either do not contribute much prominent morphology to the Creole (although the description of morphology in Creoles might be biased against the description of morphological inflection; Haspelmath and Michaelis 2017).

These findings do not entail that Creoles are just a combination of linguistic traits identical to their ancestral languages. Sometimes Creoles diverge from their ancestors, which is unsurprising as they have been in use for hundreds of years which leaves room for differences due to language change. However, the differences between Creoles and their ancestral languages were not systematic—in particular, the authors found no evidence of any systematically emerging Creole trait that could not be accounted for by either their lexifiers or their substrates. In sum, Blasi et al. (2017) determined that the Creole 'profile' that has been discussed in the literature is little more than a historical residue of the fact that most of those languages involved a very particular combination of lexifier and substrate languages. There is no evidence that Creoles are exceptional in this regard, or that they are a proxy for some type of 'preferred' grammar brought by either human biology or some propensity towards 'efficient' grammar. Evidence is lacking as well for the historical connection between pidgins and Creoles, which might have had complementary distributions (Mufwene 2020).

In fact, this reveals how diverse and rich the group of Creoles is. Outside the Atlantic region, we can find Creoles which do not conform to the 'prototypical' structural profile of Creole languages (see Ansaldo and Meyerhoff 2020 on the diversity of Creoles). For instance, although its lexifier is a highly isolating variety of Malay, Sri Lanka Malay developed into a morphologically agglutinative Creole with fusional tendencies, under the influence of Sinhala and Lankan Tamil, the dominant languages of Sri Lanka (Ansaldo 2009, 2010). This direction of change is the exact opposite of the commonly observed pattern in Atlantic Creoles. More generally, what is evident from a case like Sri Lanka Malay is that the structural outcome of language contact is heavily influenced by the nature of the grammars in contact language change (Aboh and Ansaldo 2007).

9780192886491_Book.indb 199 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:43

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

11.4 Creole emergence from the point of view of the individuals

The history of Creoles reveals that they broadly do what languages classified as non-Creole do—continuing the linguistic structures present in previous generations. Nevertheless, it remains to be explained how language learning, social interaction, and individual differences play out in Creole emergence. Crucially, in most but a few cases, detailed information on the early stages of a Creole are absent (Sranan Tongo is a potential exception), so the role and nature of the individuals involved is often extrapolated from the structure of the resulting languages.

In contrast to the notion that children are the sole responsible group for Creole emergence, adults play an important role. Crucially, their effect on the Creole can be understood through the lens of general observations stemming from second language (L2) acquisition (SLA) (Lefebvre 1998; Chaudenson 2001; Siegel 2006, 2008; Clements 2009). In particular, it has been pointed out that patterns in non-Creole SLA often match those reflected by 'substrate transfer', i.e. the process through which the substrates condition and bias the patterns in the lexifier to be acquired (Mufwene 1990). For instance, both African Portuguese-lexified Creoles and L2 African Portuguese involved similar languages: West African languages and Portuguese. In both cases, there are a number of salient morphosyntactic features absent in Portuguese (including the lack of a copula verb, the use of an all-purpose preposition, and the use of temporal adverbials as tense markers with no corresponding verbal inflection; Clements 2009). Yet, findings on SLA do not seem to always translate to Creoles, which shows again that Creoles are more than simple L2 varieties of their lexifiers (Baker 1994; Muysken 2001). For instance, many Southeast Asian Creoles have exapted a word meaning 'already' as a way of marking when an event is conceptualized as a finalized whole (referred to as perfective aspect):

(1) Singlish (English-lexified)
He eat *already*
'He has eaten.'

(2) Macanese (Portuguese-lexified) (Ferreira 1996)
Ilôtro *já* faze
they already do
'They did.'

(3) Papiá Kristang (Portuguese-lexified) (Baxter 2013)
Eli *ja* bai mar
s/he already go sea
'S/he went fishing.'

9780192886491_Book.indb 200 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:43

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

However, that very salient pattern is unobserved in L2 English of L1 speakers of Cantonese (and other structurally similar languages) (Szeto et al. 2017, 2019). One proposed explanation of this difference is that L1 acquisition can assign new grammatical functions to lexifier material, unlike typical L2 acquisition (Kouwenberg 2006).

Besides general observations from L2 and L1 acquisition, accounting for Creole genesis requires accounting for the diversity in language experiences at their onset. In particular, it needs to be explained why some of the many possible linguistic features in these multilingual setups ended up prevailing in the Creoles and others not. One framework that attempts to explain this is that of Mufwene's (2002), which posits that all language users contribute to a pool of features from which (i) each learner 'selects' a specific subset to create their own idiolect (i.e. an individual's unique variety of a language), and (ii) a speaker might choose new variants as they accommodate their interlocutors during interaction. Recently, it has been suggested that the successful acquisition and combination of variants might depend on the similarity between already acquired and newly encountered variants (Baptista 2020; Labotka et al. 2023).

11.5 Parallels between Creoles and sign languages in language evolution research

This section reflects on how Creoles and sign languages are invoked in language evolution arguments and is motivated by parallels that have been drawn between the two (Fischer 1978; Ladd and Edwards 1982; Kegl 2008; Adone 2012; Gil 2014; Bakker 2015; McWhorter 2022; Bisnath et al. under review). In particular, both Creoles and sign languages have been proposed as windows onto protolinguistic stages of language evolution on the basis of two arguments: (1) they represent a transmission 'break' in relation to other languages; and (2) they are 'young' languages.

The case for a transmission break at the core of Creoles and sign languages is weak at best. As discussed in section 11.3, Creole languages show that language transmission is remarkably robust in Creole genesis and that it is possible to trace the grammar of Creoles to the languages that coexisted and precede them (Blasi et al. 2017). The case for the emergence of sign languages seems stronger *a priori*, given that there is often a modality discontinuity. The crucial assumption is that they have developed in a linguistic vacuum as they emerged in a situation in which deaf individuals did not have access to sign language[1] (Goldin-Meadow and Yang 2017). However, this ignores forms of language that *are* available to deaf children and which are known to influence sign language structure and use, for example, homesign, spoken language mouth patterns, and written

[1] Note that there is a broad consensus that deaf children have a right to access a sign language (Jokinen 2000; Humphries et al. 2014, and references within).

9780192886491_Book.indb 201 /12_first_proofs/first_pr…/xml_for_typesetting 27-Mar-25 23:02:43

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

language (Sutton-Spence 1999; Padden and Gunsauls 2003; Koulidobrova and Pichler 2021; Bauer and Kyuseva 2022; Bisnath 2024).

The case for the 'youth' of Creoles and sign languages is also derived from the assumption there is a transmission break. One of the purported consequences of this assumed linguistic youth is that these languages are 'simpler', as language complexity is assumed to only emerge over the course of generations and with the need for more functions than just immediate communication (McWhorter 2001; Parkvall 2008; Bakker et al. 2011; Goldshtein 2017). The maturity of a language is measured as growth of structural complexity (cf. DeGraff 2001; Bisnath et al. under review). In the case of Creole languages, this notion has been repeatedly rejected (DeGraff 2005; Aboh 2016; Murawaki 2016; Blasi et al. 2017; Szeto et al. 2019; Meakins 2022). As discussed in section 11.3, Creole languages might have a more analytic flavour to them because many of the substrates involved in their emergence were analytic as well (Szeto et al. 2019). In spite of that, this idea has been transferred into sign linguistics, leading to predictions that sign languages as a group should be similarly 'simple' (Aronoff et al. 2005).

Finally, Creole and sign languages have in common the fact that their users have not been historically treated with the same dignity as other human groups. They have been dehumanized, which has direct consequences to their study (Ladd and Edwards 1982; Baynton 1996; DeGraff 2001, 2005, 2020; Krausneker 2015; Henner and Robinson 2021). In the study of Creoles, this is typically identified as Creole Exceptionalism, which has been shown to be undergirded by racist colonial ideology (DeGraff 2005). Similarly, the study of sign languages and their users was affected by audism (discrimination against deaf people) (Baynton 1996; Eckert and Rowley 2013; Krausneker 2015; Henner and Robinson 2021). Sometimes, multiple dimensions of minoritization interact within these groups. The best example comes from an assumed developmental cline in sign language development, shown in Figure 11.1 (cf. Occhino and Wilcox 2017; Kusters et al. 2020; Hou and de Vos 2022). Here, certain kinds of sign language groupings, e.g. village sign language, are framed as particularly relevant to language evolution (Botha 2007; Benítez-Burraco 2015; Jackendoff and Wittenberg 2017). However, this does not reflect the critique of sign language researchers who find the evidence for separation between homesign, village sign language, etc., weak (Botha 2007; Benítez-Burraco 2015; Jackendoff and Wittenberg 2017; cf. Kusters et al. 2020, Bisnath et al. under review). This separation may reflect global power inequalities as village sign languages tend to be found in the global south (Braithwaite 2020).

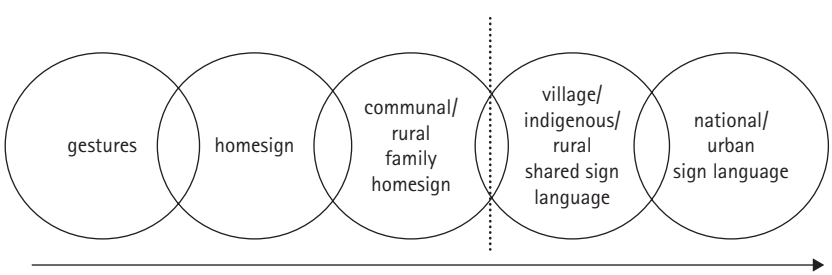


FIGURE 11.1 Developmental cline.

Source: adapted from Kusters et al. (2020: 12).

9780192886491_Book.indb 202 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:43

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

11.6 The contribution of Creole languages to the field of Language Evolution

The bioprogram hypothesis established Creole languages as a shortcut to the early stages of language, irrevocably lost to time. As we have discussed in sections 11.3 and 11.5, this and other ideas have not survived empirical scrutiny, and have likely persisted due to lack of interdisciplinary exchange and/or deeply seated beliefs about Creole languages and their users. Creoles are not mysterious in their origin: they continue (with modifications) the structure of the languages that the previous generations of users had access to—as is the case with non-Creole languages. Yet we should not throw the baby out with the bathwater, as Creoles do provide us with insightful information about the processes of language transmission and language emergence.

To start with, Creoles show how different language transmission can be from the most well-studied setups of WEIRD (White, Educated, Industrialized, Rich, Democratic) societies (Henrich et al. 2022). The idea that large amounts of directed and tailored input are a necessity for the development and the acquisition of language is at odds with the exuberant multilingualism and the lack of formal education that characterize many Creole setups.

Creole emergence also reveals intriguing patterns of association between different components of language transmission in multilingual contexts (Mufwene 2002)—for instance, word order patterns in Creoles largely follow that of the lexifiers (Blasi et al. 2017). It has been argued that the dynamics of language in multilingual societies is the appropriate model to consider when making inferences about prehistoric stages of language diversity (Evans 2017).

Finally, any solid model of the evolution of language needs to draw from the set of cognitive and communicative phenomena displayed by humans when learning and creating the wealth of language diversity we have access to. A handful of languages, most prominently English, have dominated the intuitions and targets of explanation when it comes to language evolution, mimicking what happens elsewhere in cognitive science (Blasi et al. 2022; Majid 2023). This situation is particularly intriguing in the case of Creole languages: they are not a marginal group of languages, as they are found in virtually all continents and they are used by hundreds of millions around the world. We hope that, in the future, Creole languages will be regarded in the field not as ahistorical reflections of hypothetical protolanguages, but as evidence that humans can and do create languages in the most diverse situations imaginable, a living proof of human ingenuity.

References

Aboh, E. O. (2016). 'Creole distinctiveness: A dead end', *Journal of Pidgin and Creole Languages* 31(2): 400–418.

9780192886491_Book.indb 203 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:43

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Aboh, E. O. and Ansaldo, U. (2007). 'The role of typology in language creation', in U. Ansaldo, S. Matthews, and L. Lim (eds.), *Deconstructing Creole (Typological Studies in Language)*, 39–66. Amsterdam: John Benjamins.

Adone, D. (2012). 'Language emergence and creolisation', in R. Pfau, M. Steinbach, and B. Woll (eds.), *Sign Language: An International Handbook* (Handbücher Zur Sprach- Und Kommunikationswissenschaft; Handbooks of Linguistics and Communication Science Bd. 37 = Bd. 37), 862–888. Berlin, Boston: De Gruyter Mouton.

Ansaldo, U. (2009). *Contact Languages: Ecology and Evolution in Asia*. New York: Cambridge University Press.

Ansaldo, U. (2010). 'Contact and Asian varieties of English', in R. Hickey (eds.), *The Handbook of Language Contact*, 498–517. New York: Wiley Online Library.

Ansaldo, U. and Meyerhoff, M. (eds.). (2020). *The Routledge Handbook of Pidgin and Creole Languages*. 1st edn, https://doi.org/10.4324/9781003107224.

Ansaldo, U. and Szeto, P. Y. (2020). 'Pidgin and Creole ecology and evolution', in U. Ansaldo and M. Meyerhoff (eds.), *The Routledge Handbook of Pidgin and Creole Languages*, 504–519. London: Routledge.

Arends, J. (1995). 'Demographic factors in the formation of Sranan', *Creole Language Library* 13: 233–285.

Aronoff, M., Meir, I., and Sandler, W. (2005). 'The paradox of sign language morphology', *Language* 81(2): 301–344, https://doi.org/10.1353/lan.2005.0043.

Baker, P. (1994). 'Creativity in creole genesis', in D. Adone and I. Plag (eds.), *Creolization and Language Change*, 65–84. Berlin, New York: De Gruyter, https://doi.org/10.1515/9783111339801.65.

Baker, P. and Kriegel, S. (2013). 'Mauritian Creole', in S. Michaelis, P. Maurer, M. Haspelmath, and M. Huber (eds.), *The Survey of Pidgin and Creole Languages*, vol. 2, 250–260. Oxford: Oxford University Press.

Bakker, P. (2015). 'Creoles, creole studies and sign languages', *Journal of Pidgin and Creole Languages* 30(2): 357–369, https://doi.org/10.1075/jpcl.30.2.08bak.

Bakker, P., Daval-Markussen, A., Parkvall, M., and Plag, I. (2011). 'Creoles are typologically distinct from non-creoles', *Journal of Pidgin and Creole Languages* 26(1): 5–42.

Bancu, A., Peltier, J., Bisnath, F., et al. 2024. 'On "revitalizing" attitudes towards Creole languages', in H. Anne, C. Mallinson, and M. Bucholtz (eds.), *Decolonizing Linguistics*, 293–316. New York: Oxford University Press.

Baptista, M. (2020). 'Competition, selection, and the role of congruence in creole genesis and development'. *Language* 96(1): 160–199.

Bauer, A. and Kyuseva, M. (2022). 'New insights into Mouthings: Evidence from a corpus-based study of Russian sign language', *Frontiers in Psychology* 12: 6496.

Baxter, A. N. (2013). 'Papiá Kristang structure dataset', in S. M. Michaelis, P. Maurer, M. Haspelmath, and M. Huber (eds.), *Atlas of Pidgin and Creole Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology, https://apics-online.info/contributions/42.

Baynton, D. C. (1996). *Forbidden Signs: American Culture and the Campaign Against Sign Language*. Chicago: University of Chicago Press.

Benítez-Burraco, A. (2015). 'A biolinguistic approach to sign languages', in M. Marschark and P. E. Spencer (eds.), *The Oxford Handbook of Deaf Studies in Language*, 247–263. New York: Oxford University Press, https://doi.org/10.1093/oxfordhb/9780190241414.001.0001.

9780192886491_Book.indb 204 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:43

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Bickerton, D. (1984). 'The language bioprogram hypothesis', *Behavioral and Brain Sciences* 7(2): 173–188.

Bisnath, F. (2024). 'Mouthing constructions in 37 signed languages: Typology, ecology and ideology', *Journal of Language Contact* 16(4): 565–614, https://doi.org/10.31234/osf.io/rxm7w.

Bisnath, F., Jaraisy, M., Lutzenberger, H., et al. (Under review). 'Deconstructing notions of morphological "complexity": Lessons from signed and spoken languages'.

Blasi, D. E., Michaelis, S. M., and Haspelmath, M. (2017). 'Grammars are robustly transmitted even during the emergence of creole languages', *Nature Human Behaviour* 1(10): 723–729, https://doi.org/10.1038/s41562-017-0192-4.

Blasi, D. E., Henrich, J., Adamou, E., et al. (2022). 'Over-reliance on English hinders cognitive science', *Trends in Cognitive Sciences* 26(12): 1153–1170, https://doi.org/10.1016/j.tics.2022.09.015.

Botha, R. (2007). 'On homesign systems as a potential window on language evolution', *Language & Communication* 27(1): 41–53.

Bouchard, M.-E. (2022). 'The vitality of Angolar: A study of attitudes on São Tomé Island', *Journal of Pidgin and Creole Languages* 37(1): 160–188, https://doi.org/10.1075/jpcl.00089.bou.

Braithwaite, B. (2020). 'Ideologies of linguistic research on small sign languages in the global South: A Caribbean perspective', *Language & Communication* 74: 182–194, https://doi.org/10.1016/j.langcom.2020.06.009.

Chaudenson, R. (2001). *Creolization of Language and Culture*. London: Routledge.

Clements, J. C. (2009). *The Linguistic Legacy of Spanish and Portuguese: Colonial Expansion and Language Change*. 1st edn. New York: Cambridge University Press, https://doi.org/10.1017/CBO9780511576171.

Danchev, A. (1997). 'The Middle English creolization hypothesis revisited', *Trends In Linguistics Studies and Monographs* 103: 79–108.

DeGraff, M. (2001). 'On the origin of creoles: A Cartesian critique of neo-Darwinian linguistics', *Linguistic Typology* 5(2/3): 213–310.

DeGraff, M. (2005). 'Linguists' most dangerous myth: The fallacy of Creole Exceptionalism', *Language in Society* 34(4): 533–591, https://doi.org/10.1017/S0047404505050207.

DeGraff, M. (2020). 'Toward racial justice in linguistics: The case of Creole studies (Response to Charity Hudley et al.)', *Language* 96(4): e292–e306.

DeLancey, S. (2014). 'Creolization in the divergence of the Tibeto-Burman languages', in T. Owen-Smith and N. Hill (eds.), *Trans-Himalayan Linguistics*, 41–70. Berlin, Boston: De Gruyter, https://doi.org/10.1515/9783110310832.41.

Eckert, R. C. and Rowley, A. J. (2013). 'Audism: A theory and practice of audiocentric privilege', *Humanity & Society* 37(2): 101–130, https://doi.org/10.1177/0160597613481731.

Eriksen, T. H. (2007). 'Creolization in anthropological theory and in Mauritius', in C. Stewart (ed.), *Creolization: History, Ethnography, Theory*, 153–177. New York: Routledge.

Evans, N. (2017). 'Did language evolve in multilingual settings?', *Biology & Philosophy* 32(6): 905–933, https://doi.org/10.1007/s10539-018-9609-3.

Ferreira, J. dos S. (1996). *Papiaçám di Macau.* Macau: Fundação Macau.

Fischer, S. D. (1978). 'Sign languages and creoles', in P. Siple (ed.), *Understanding Language Through Sign Language Research*, 309–331. New York: Academic Press.

Gil, D. (2014). 'Sign languages, Creoles and the development of predication', in *Measuring Grammatical Complexity*, 37–64. Oxford: Oxford University Press.

9780192886491_Book.indb 205 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:43

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Goldin-Meadow, S. and Yang, C. (2017). 'Statistical evidence that a child can create a combinatorial linguistic system without external linguistic input: Implications for language evolution', *Neuroscience & Biobehavioral Reviews* 81: 150–157.

Goldshtein, Y. (2017). 'The simple emerging from the complex', in P. Bakker, F. Borchsensius, C. Levisen, and E. M. Sippola (eds.), *Creole Studies–Phylogenetic Approaches*, 193–217. Amsterdam: John Benjamins Publishing Company.

Haspelmath, M. and Michaelis, S. M. (2017). 'Analytic and synthetic: Typological change in varieties of European languages', in I. Buchstaller and B. Siebenhaar (eds.), *Studies in Language Variation*, vol. 19, 3–22. Amsterdam: John Benjamins Publishing Company, https://doi.org/10.1075/silv.19.01has.

Henner, J. and Robinson, O. (2021). 'Unsettling languages, unruly bodyminds: Imaging a Crip linguistics', https://doi.org/10.31234/osf.io/7bzaw.

Henrich, J., Blasi, D. E., Curtin, C. M., et al. (2022). 'A cultural species and its cognitive phenotypes: Implications for philosophy', *Review of Philosophy and Psychology* 14: 349–386, https://doi.org/10.1007/s13164-021-00612-y.

Hou, L. and de Vos, C. (2022). 'Classifications and typologies: Labeling sign languages and signing communities', *Journal of Sociolinguistics* 26(1): 118–125.

Humphries, T., Kushalnagar, P., Mathur, G., et al. (2014). 'Ensuring language acquisition for deaf children: What linguists can do', *Language* 90(2): e31–e52.

Jackendoff, R. and Wittenberg, E. (2017). 'Linear grammar as a possible stepping-stone in the evolution of language', *Psychonomic Bulletin & Review* 24(1): 219–224, https://doi.org/10.3758/s13423-016-1073-y.

Jokinen, M. (2000). 'The linguistic human rights of sign language users', in R. Philipson (ed.), *Rights to Language*, 207–217. 1st edn. New York: Routledge, https://doi.org/10.4324/9781410605603.

Kegl, J. (2008). 'The case of signed languages in the context of pidgin and creole studies', in S. Kouwenberg and J. Singler (eds.), *The Handbook of Pidgin and Creole Studies*, 489–511. West Sussex: John Wiley.

Koulidobrova, E. and Chen Pichler, D. (2021). 'Cripping the "delay": Multilingualism-related consequences of re-labeling language deprivation systems', *Frontiers in Communication* 6: 751457.

Kouwenberg, S. (2006). 'LI transfer and the cut-off point for L2 acquisition processes in Creole formation', in C. Lefebvre, L. White, and C. Jourdan (eds.), *Language Acquisition and Language Disorders*, vol. 42, 205–219. Amsterdam: John Benjamins Publishing Company, https://doi.org/10.1075/lald.42.13kou.

Krausneker, V. (2015). 'Ideologies and attitudes toward sign languages: An approximation', *Sign Language Studies* 15(4): 411–431, https://doi.org/10.1353/sls.2015.0014.

Kusters, A., Green, M., Moriarty, E., and Snoddon, K. (2020). 'Sign language ideologies: Practices and politics', in A. Kusters, M. Green, E. Moriarty, and K. Snoddon (eds.), *Sign Language Ideologies in Practice* (Sign Languages and Deaf Communities [SLDC]), vol. 12, 3–22. Berlin, Boston: De Gruyter, https://doi.org/10.1515/9781501510090-001.

Labotka, D., Sabo, E., Bonais, R., et al. (2023). 'Testing the effects of congruence in adult multilingual acquisition with implications for creole genesis', *Cognition* 235: 105387.

Ladd, P. and Edwards, V. (1982). 'British Sign Language & West Indian Creole', *Sign Language Studies* 1035(1): 101–126, https://doi.org/10.1353/sls.1982.0028.

Lee, N. H. (2021). 'South and Southeast Asia', in Ansaldo, U. and Meyerhoff, M. (eds.), *The Routledge Handbook of Pidgin and Creole Languages*, 74–87. London, New York: Routledge, Taylor & Francis Group.

9780192886491_Book.indb 206 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:43

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Lefebvre, C. (1998). *Creole Genesis and the Acquisition of Grammar: The Case for Haitian Creole:* vol. 88. Cambridge: Cambridge University Press.

Lim, L. and Ansaldo, U. (2016). *Languages in Contact*. Cambridge: Cambridge University Press.

Lupton, L. and Salmons, J. (1996). 'A re-analysis of the Creole status of American Sign Language', *Sign Language Studies* 1090(1): 80–94, https://doi.org/10.1353/sls.1996.0013.

Majid, A. (2023). 'Establishing psychological universals', *Nature Reviews Psychology* 2(4): 199–200, https://doi.org/10.1038/s44159-023-00169-w.

McWhorter, J. H. (2001). 'The world's simplest grammars are creole grammars', *Linguistic Typology* 5(2–3): 125–166, https://doi.org/10.1515/lity.2001.001.

McWhorter, J. (2002). 'What happened to English?', *Diachronica* 19(2): 217–272, https://doi.org/10.1075/dia.19.2.02wha.

McWhorter, J. (2022). 'Spoken and sign language emergence: A comparison', *Languages* 7(3): 184.

Meakins, F. (2022). 'Empiricism or imperialism: The science of Creole Exceptionalism', *Journal of Pidgin and Creole Languages* 37(1): 189–203, https://doi.org/10.1075/jpcl.00092.mea.

Michaelis, S. M. (2020). 'Avoiding bias in comparative creole studies: Stratification by lexifier and substrate', *Isogloss: Open Journal of Romance Linguistics* 6: 1, https://doi.org/10.5565/rev/isogloss.100.

Mufwene, S. S. (1990). 'Transfer and the substrate hypothesis in creolistics', *Studies in Second Language Acquisition* 12(1): 1–23.

Mufwene, S. S. (2002). 'Competition and selection in language evolution', *Selection* 3(1): 45–56, https://doi.org/10.1556/Select.3.2002.1.5.

Mufwene, S. S. (2020). 'Creoles and pidgins: Why the latter are not the ancestors of the former', in E. Adamou and Y. Matras (eds.), *The Routledge Handbook of Language Contact*, 300–324. London: Routledge.

Mühleisen, S. (2022). 'Talking about Creole: Language attitudes and public discourse in the Caribbean', *Social and Structural Aspects of Language Contact and Change* 5: 263.

Murawaki, Y. (2016). 'Statistical modeling of creole genesis', in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1329–1339.

Muysken, P. (2001). 'The origin of creole languages: The perspective of second language learning', in N. Smith and T. Veenstra (eds.), *Creole Language Library*, vol. 23, 157. Amsterdam: John Benjamins Publishing Company, https://doi.org/10.1075/cll.23.07muy.

Nero, S. J. (2014). 'De facto language education policy through teachers' attitudes and practices: A critical ethnographic study in three Jamaican schools', *Language Policy* 13: 221–242.

Occhino, C. and Wilcox, S. (2017). 'Gesture or sign? A categorization problem', *Behavioral and Brain Sciences* 40: e66.

Padden, C. A. and Gunsauls, D. C. (2003). 'How the alphabet came to be used in a sign language', *Sign Language Studies* 4(1): 10–33.

Palmié, S. (2016). 'The "C-Word" again: From colonial to postcolonial semantics', in C. Stewart (ed.), *Creolization*, 74–91. New York: Routledge.

Parkvall, M. (2008). 'The simplicity of creoles in a cross-linguistic perspective', in M. Miestamo, K. Sinnemäki, and F. Karlsson (eds.), *Language Complexity: Typology, Contact and Change* (Studies in Language Companion Series (SLCS) 94), 265–286. Amsterdam, Philadelphia: John Benjamins Publishing Company.

Pennycook, A. (2007). 'The myth of English as an international language', *Disinventing and Reconstituting Languages* 62: 90–115.

9780192886491_Book.indb 207 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:43

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Rickford, J. R. (1985). 'Standard and non-standard language attitudes in a creole continuum', in N. Wolfson and J. Manes (eds.), *Language of Inequality*, 145–160. Berlin, Boston: De Gruyter, https://doi.org/10.1515/9783110857320.145.

Rickford, J. R. (2021). 'The creole origins of African-American Vernacular English: Evidence from copula absence', *African-American English*, 169–220. London: Routledge.

Romaine, S. and Rickford, J. R. (1999). *Creole Genesis, Attitudes and Discourse* (Creole Language Library). Amsterdam, Philadelphia, PA: John Benjamins Publishing Company.

Sanogo, Y. A. and Kamei, N. (2019). 'La promotion de la recherche sur la Langue des Signes par les communautés des Sourds africains. Cas de l'Afrique de l'Ouest et de Centre francophone' 共生の文化研究= *Journal of cultural symbiosis research*. 愛知県立大学多文化共生研究所 13: 5–16.

Siegel, J. (2006). 'Links between SLA and Creole studies: Past and present', in C. Lefebvre, L. White, and C. Jourdan (eds.), *Language Acquisition and Language Disorders*, vol. 42, 15–46. Amsterdam: John Benjamins Publishing Company, https://doi.org/10.1075/lald.42.03sie.

Siegel, J. (2008). *The Emergence of Pidgin and Creole Languages* (Oxford Linguistics). Oxford, New York: Oxford University Press.

Sutton-Spence, R. (1999). 'The influence of English on British sign language', *International Journal of Bilingualism* 3(4): 363–394.

Szeto, P. Y., Matthews, S., and Yip, V. (2017). 'Multiple correspondence and typological convergence in contact-induced grammaticalization: Evidence from Cantonese-English bilingual development', *Journal of Language Contact* 10(3): 485–518.

Szeto, P. Y., Lai, J. Y., and Ansaldo, U. (2019). 'Creole typology is analytic typology', *Language Ecology* 3(1): 89–119.

Szeto, P. Y., Matthews, S., and Yip, V. (2019). 'Bilingual children as "laboratories" for studying contact outcomes: Development of perfective aspect', *Linguistics* 57(3): 693–723.

Velupillai, V. (2015). *Pidgin and Creole Languages. An Introduction*. Amsterdam, Philadelphia: John Benjamins.

Wassink, A. B. (1999). 'Historic low prestige and seeds of change: Attitudes toward Jamaican Creole', *Language in Society* 28(01): https://doi.org/10.1017/S0047404599001037.

Woodward, J. (1978). 'Historical bases of American Sign Language', in P. Siple (ed.), *Understanding Language Through Sign Language Research*, 334–348. New York: Academic Press, http://www.cslds.org/v4/resources/InternalAdmin/publicationB/16/_Woodward_1978_HistoricalBasesOfASL_.pdf. (30 April 2023).

9780192886491_Book.indb 208 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:44

CHAPTER 12

PALAEOGENOMICS

A window into the genetic basis of derived traits in Homo sapiens

JUAN MORIANO, CEDRIC BOECKX, AND MARTIN KUHLWILM

The absence of fossilized brain tissue in the archaeological record significantly hinders inferences about the evolution of the neurobiological underpinnings of defining cognitive traits of our species, such as language. In a landmark publication in 1997, mitochondrial DNA was retrieved from a 0.4 g sample of the right humerus bone of a Neanderthal fossil in the Feldhofer cave (Neander Valley, Germany) (Krings et al. 1997). Almost fifteen years later, a new hominin species was discovered solely through palaeogenomic analysis from extracts of a finger bone found in the Denisova cave in Siberia (Reich et al. 2010). What was once described as the emerging field of molecular archaeology (Pääbo et al. 1989) is now a burgeoning interdisciplinary field dramatically expanding the range of questions researchers can ask in the quest to understand the human condition (Pääbo 2014). The retrieval of ancient DNA, defined as extremely short DNA fragments preserved in the fossil record and extracted from material such as bones, teeth, skin, hair, or sediments (Orlando et al. 2021) has enabled direct comparisons of the genome of our species to that of our closest extinct relatives, the Neanderthals and Denisovans. This is particularly interesting, since our closest living relatives—chimpanzees and bonobos—share a common ancestor with the *Homo* lineage six to ten million years ago (Besenbacher et al. 2019), which is an order of magnitude more than the split of modern humans from these extinct hominin lineages. The deep divergence means that there is a wealth of genetic and phenotypic differences to the great apes (Varki and Altheide 2005). While most of these differences are shared among hominins, their genomes allow a unique view into 'modern' human-specific features that are not shared, but instead show the ancestral state in the archaic lineages. Hence, palaeogenomics provides a unique entry point to decipher the genetic basis of what makes us human. However, connecting mutations in a DNA sequence to specific molecular, cellular, and higher

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

phenotypic levels remains a fundamental problem in biology, and one that requires insights from multiple disciplines, with genetics being only one of them. Additionally, as particularly evident in the field of the neurobiology of language, efforts attempting to resolve the biological underpinnings of traits such as the human faculty of language require the right level of granularity across research domains and explanatory linking hypotheses (Poeppel 2012).

A paradigmatic case, highlighting the potential of palaeogenomic research, is provided by the study of the forkhead box protein P2, *FOXP2*. This transcription factor, unequivocally associated to key aspects of (spoken) language, was discovered after the characterization of a three-generation family where around half of the members, with severe linguistic impairments, were found to carry a heterozygous missense mutation in the *FOXP2* gene (Lai et al. 2001). Experimental research in a variety of model systems has significantly advanced our understanding of the functional roles of FOXP2. The introduction of the arginine-to-histidine substitution present in the affected KE family members in the homologue gene in mice was reported to impair ultrasonic vocalizations with less complex sequences (Chabout et al. 2016), whereas its knockdown in zebra finches impacts dopaminergic signalling and song variability (Murugan et al. 2013). Additionally, brain imaging studies have revealed that affected members of the KE family present abnormalities in brain anatomy beyond the neocortex, implicating subcortical structures such as the caudate nucleus or the putamen (Watkins et al. 2002; Liégeois et al. 2003). In light of its possible key roles for the emergence of language in our species, studies focused on the evolution of the FOXP2 gene, and found that despite its high sequence conservation, the human gene harbours two non-synonymous mutations (both in exon 7) that emerged after the last common ancestor with the chimpanzees (Enard et al. 2002; Zhang et al. 2002). Subsequent palaeogenomics analyses on ancient DNA from Neanderthal remains found at El Sidrón Cave in Asturias (Spain) revealed that these two amino acid changes were shared with the Neanderthals, concluding that most likely those substitutions were present in the last common ancestor of *Homo sapiens* and the Neanderthals (Krause et al. 2007). It is expected that most of the fixed or almost fixed genetic differences arose through drift and likely do not exert any differential functional roles, but those that were under selective pressures in our lineage not only affected coding regions but also impacted the regulatory landscape that controls gene expression in a spatiotemporal precise manner. Adding evidence to the evolving story of *FOXP2* (Fisher 2019), a derived variant in the *Homo sapiens* lineage located in the intron 8 of *FOXP2* was shown to cause differential transcription factor binding of POU3F2 (Maricic et al. 2013). The functional consequences of this change remain unclear, and a recent study found that the intronic region is transcribed and reported no evidence of positive selection in the locus (Atkinson et al. 2018). As revealed by the FOXP2 case, ancient DNA research can illuminate inquiries into what sets our species apart from our closest extinct relatives going beyond comparisons with non-human primates, and clearly exemplifies the need to bring together diverse analytical and experimental approaches to explore the questions about our species' cognitive abilities. But just like there is no gene for language, there is no gene for language evolution, and

9780192886491_Book.indb 210 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:44

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

no single mutation will be able to explain changes to our language-ready brain (de Boer et al. 2020).

Evidence for the functional consequences of derived variants in *Homo sapiens*, when compared to our closest relatives, implicate diverse systems such as the skin, the immune system, the metabolism, or the skeleton. Likewise, some of these variants are expected to have impacted the nervous system. In this chapter, we focus on the use of palaeogenomics to reconstruct ancient phenotypes of relevance for human brain evolution, and on approaches to identify and prioritize candidates for experimental interrogation. The emergence of the so-called modern human condition, and the evolution of the 'modern' faculty of language in particular, was plausibly accompanied by the gradual accumulation of genetic mutations rather than the result from a single mutation, and the efforts now propelled by the field of palaeogenomics urge a comprehensive characterization of the genetic mutations that were under positive selection in our lineage. Of note, the quantitative revolution in the field of developmental neuroscience (Briscoe and Marín 2020) closely aligns to the efforts to overcome the genotype-to-phenotype problem presented above, and a final section is dedicated to examples of how concrete hypotheses facing this mapping problem, to explain the evolution of modern human traits, can be experimentally tested.

12.1 The resurrection of ancient *Homo* genomes

The degradation and chemical modification of ancient DNA material, along with the contamination from microbial and human sources, pose severe challenges for the reconstruction of the genomes of ancient populations. For instance, early insights obtained from high-throughput sequencing of nuclear Neanderthal genomes were later shown to be significantly contaminated by non-endogenous DNA (Green et al. 2006; Wall and Kim 2007). Advancements in the purification, amplification, and high-throughput sequencing of ancient DNA now allows us to confidently reconstruct at high coverage the genomes of our closest relatives with a quality comparable to that obtained from present-day humans. Comparative genomics using ancient *Homo* genomes has been fostered likewise by a wider, more diverse characterization of genomes from present-day human populations across continents, thanks to efforts like the 1000 Genomes Project (Auton et al. 2015) or the Simons Genome Diversity Project (Mallick et al. 2016), as well as by the genomes of other primate species (Kuderna et al. 2023). At present, four high-coverage archaic genomes have been obtained at exceptional quality from remains belonging to both the Neanderthals and Denisovan species. In 2012, Meyer et al. (2012) sequenced the genome of a female Denisovan individual utilizing DNA extracts from a finger phalanx from the Denisova Cave in the Altai Mountains. The genome of this Denisovan individual was obtained at a thirty-fold

9780192886491_Book.indb 211 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:44

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

coverage by employing a single-stranded DNA library preparation, which significantly enhanced the final sequencing yield compared to previous efforts (Meyer et al. 2012). Two years later, the genome of a female Neanderthal individual—also from the Denisova cave—was reconstructed at a fifty-two-fold coverage (Prüfer et al. 2014). This first high-quality genome from a Neanderthal provided estimates of a population split time of Neanderthal/Denisovans from the *Homo sapiens* lineage between 550,000 and 765,000 years ago, and aided in the depiction of the complex evolutionary history of our ancestors, estimating the level of Neanderthal introgressed DNA material in current modern human populations at around 1.5–2.1 per cent, as well as providing evidence for Neanderthal introgression into Denisovans (Prüfer et al. 2014). The first two high-coverage genomes led to the first comprehensive catalogue of single nucleotide variants and short insertions and deletions that distinguish our species since our divergence from the Neanderthal/Denisovan lineage. As a result, a small set of proteins (less than one hundred) were identified carrying fixed amino acid substitutions along with a set of regulatory variants derived in our species where the Neanderthals/Denisovans carry the ancestral allele found in non-human primates. Proteins carrying non-synonymous substitutions derived in our lineage were found related to axonal and dendritic growth and synaptic transmission and, prominently, an overrepresentation of genes relevant for neural progenitor cell behaviour was reported (Prüfer et al. 2014). Two additional Neanderthal genomes have been reconstructed over the past decade. A second genome of a female Neanderthal was sequenced at thirty-fold coverage from remains at the Vindija cave (Prüfer et al. 2017) and a third Neanderthal genome was uncovered from remains of Chagyrskaya Cave in the Altai Mountains, sequenced at a twenty-seven-fold genomic coverage (Mafessoni et al. 2020). This is supplemented by several low-to-medium coverage Neanderthal and Denisovan genomes (Slon et al. 2017; Hajdinjak et al. 2018; Skov et al. 2022), which may be used to gain additional support for the status of specific changes in the archaic lineages. Despite the few genomes currently available, the three Neanderthal genomes at high quality enabled us to more confidently elaborate a catalogue of derived genetic variants in the Neanderthal lineage, revealing that genes expressed in the striatum rather than other brain genes show derived features in the Neanderthal lineage, while some of the fixed changes within genes expressed in the striatum might have been under negative selection in anatomically modern humans (Mafessoni et al. 2020).

12.1.1 A note on low coverage genomes

While the sequencing of genomes from extinct *Homo* species at high quality represents a turning point in the field of human evolution research, the importance of genomic data from ancient individuals, even at low coverage, cannot be underestimated. This is perhaps best exemplified by the first drafts of the genomes of both the Neanderthals and the Denisovans at a coverage below two-fold. A milestone of ancient DNA in human evolution research was achieved in 2010, when Green et al. presented a first draft of

9780192886491_Book.indb 212 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:44

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

a Neanderthal genome from three different individuals at the Vindija cave (Croatia), based on a total of four billion nucleotides yielding a low coverage of 1.3-fold. These first analyses proposed as the most parsimonious explanation gene flow from Neanderthals to non-African modern human populations, based on Neanderthals sharing more alleles with non-African individuals than African individuals. Also, candidate genes under putative positive selection were unearthed and linked to cognitive aspects of our species. Furthermore, increasing the geographical and temporal resolution of ancient genomes will be pivotal to obtaining a more faithful representation of the social structure and history of extinct *Homo* populations. Hajdinjak et al. (2018) used sodium hypochlorite treatment on bone and tooth powder to drastically increase the recovery of ancient DNA from samples with low endogenous DNA content or substantial contamination. Applying this strategy on five different samples of late Neanderthals from fossils distributed across western Eurasia, Hajdinjak et al. (2018) obtained genomes with an average coverage between 1-fold and 2.7-fold. Comparative analyses, including high-quality Neanderthal genomes, Hajdinjak et al. (2018) estimated split times within Neanderthal population structure and found evidence for a population turnover during the last period of Neanderthals presence in western Eurasia. In addition, low-to-medium coverage Neanderthal genomes have further revealed familial relationships in Neanderthal communities (e.g. a father–daughter pair) as well as allowing the inference of the (small) size of such communities (Skov et al. 2022), altogether providing unprecedented insights into the complex social life of our closest relatives.

12.2 On catalogues and prioritization of variants to study the evolution of the modern human brain

Palaeogenomic analyses have revealed that only a small set of proteins, fewer than one hundred, carry fixed non-synonymous substitutions derived in *Homo sapiens*. With increasing efforts to capture the full panoply of genomic diversity in present-day populations (Auton et al. 2015; Mallick et al. 2016), the interrogation of nearly fixed substitutions in modern human populations opens up further possibilities. Examining extant variation across present-day human genomes is required to uncover nearly fixed derived variants in *Homo sapiens* that have been associated with phenotypes of clinical relevance, or to detect haplotypes from past interbreeding events and that were retained due to its adaptive value. Kuhlwilm and Boeckx (2019) stressed the relevance of such an approach revising the list of genes with protein-altering fixed changes, as well as extending the set of proteins with non-synonymous substitutions to 571 carrying high-frequency substitutions present in at least 90 per cent of present-day humans. Further exploiting the opportunities offered by the investigation of extant human genetic variation, Andirkó et al. (2022) generated an atlas with age estimates of high-frequency

9780192886491_Book.indb 213 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:44

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

alleles derived in *Homo sapiens*, which enabled a temporal resolution in which to test hypotheses on the complex evolutionary history of our species. In addition to non-synonymous mutations, the number of fixed or nearly fixed substitutions falling within non-coding regions largely exceeds protein-altering mutations, as expected. Even though a large proportion of these non-coding regions likely does not harbour any function, those that do might be relevant to explain the genetic basis of traits of evolutionary relevance, as already suggested decades ago by King and Wilson (1975) in the comparison with chimpanzees. A significant challenge in the palaeogenomic field is to narrow down the list of variants which were not random mutations rising to high frequency due to genetic drift.

12.2.1 Some lessons from archaic genomes: The impact of introgressed variants

As highlighted above, the first incursions into the genomes of extinct *Homo* species revealed admixture events between our species and the Neanderthals and Denisovans. Estimates of ancient DNA introgression present in modern human populations outside Africa are usually around 2 per cent content of Neanderthal DNA per individual (Bergström et al. 2021). Gene flow from Denisovan populations has been detected in present-day individuals from Oceania, Southeast Asia, and to a much lesser extent in East and South Asian and Native American individuals (Bergström et al. 2021). While most introgressed variants are predicted to be neutral, some might have been under positive or negative selection in our lineage (Fontsere et al. 2019). The interrogation of some of these variants have already provided evidence for effects on different aspects of modern human biology, for instance skin pigmentation, immunity, or metabolism (Reilly et al. 2022). And the functional dissection of some of these introgressed variants have uncovered prominent candidates with likely functional impacts, such as the sodium channel Nav1.7 on pain processing (Zeberg et al. 2020a), the progesterone receptor PGR and fertility (Zeberg et al. 2020b), or the glutathione reductase GR and inflammation (Coppo et al. 2022). Notably, as far as our focus—the brain—is concerned, a comprehensive analysis on more than fifty tissues and 200 individuals found a significant downregulation of Neanderthal alleles in the brain, and more prominently, in the cerebellum and the basal ganglia, among an estimated total of a quarter of introgressed variants predicted to impact gene expression regulation (McCoy et al. 2017).

A powerful avenue of research to address the aforementioned genotype-to-phenotype mapping problem builds on biomedical databases that gather large-scale neuroimaging and genomics data, enhancing genome-wide association studies (GWAS) that connect variants to neuroanatomical traits in a statistically significant manner. In a landmark study, Alagöz et al. (2022) took advantage of neuroimaging genomic data from around 30,000 participants (of European ancestry) from the UK Biobank to identify polymorphic sites previously annotated as of relevance for *Homo sapiens* brain evolution,

9780192886491_Book.indb 214 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:44

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

with a focus on heritability in specific measures of cortical area size and white matter connectivity. Variants within human-gained enhancers (compared to non-human primates) show enriched heritability in specific cortical areas, prominently Broca's area; additionally, variants within deserts of introgression (see section 12.2.2 below) showed a depletion of heritability in the uncinate fasciculus (Alagöz et al. 2022), although such results remain difficult to interpret functionally. In line with the expectation that *Homo* species-specific differences in brain organization affect regions beyond neocortical areas, genes within the deserts of introgression were found overrepresented in subcortical areas such as the striatum (Vernot et al. 2016).

12.2.2 Signatures of positive selection on the human lineage and negative selection against archaic ancestry

Since the split from the Neanderthal/Denisovan lineage, some variants reached high frequency or even fixation in our lineage and can be found in unusually long genomic regions, suggesting instances of positive selection (Racimo 2016; Peyrégne et al. 2017). For example, a set of 314 genomic regions was identified as putative positively selected regions in the *Homo sapiens* lineage, and was found enriched in regulatory regions annotated as enhancers and promoters (Peyrégne et al. 2017). Some of these regions overlap protein-coding genes carrying fixed amino acid substitutions; this is the case of *ADSL*, where evidence of functional consequence has been found (see section 12.3.1 below). Of note, only a dozen genes are present at the intersection of deserts of introgression and putative positively selected regions (Buisan et al. 2022). Some show a distinctive gene expression profile in brain regions beyond the neocortex (Buisan et al. 2022); for instance, Ca^{2+}-dependent activator protein CADPS2, which carries several fixed mutations that might impact its expression, shows a pronounced increase in expression around birth and infancy stages, peaking into adulthood in the cerebellum (Buisan et al. 2022). Experimental manipulation in mice has been found to affect cerebellar development and induce autistic-like behaviours (Sadakata et al. 2007, 2017); given the prominence of the cerebellum in the derived globular profile of modern human brains, CADPS2 stands out as a promising candidate for future experimental testing.

Complementary to the efforts to study the presence of Neanderthal ancestry in modern human genomes, genomic regions that are depleted of Neanderthal ancestry are potentially relevant in the quest to reveal the biological underpinnings of derived traits in our species. Indeed, introgressed variants are likely under negative selection in promoters and non-coding conserved genomic regions (Petr et al. 2019) and selection against Neanderthal variants possibly occurred soon after the admixture event (Hajdinjak et al. 2021; Veller et al. 2023). A recent study identified uniquely large genomic regions significantly depleted of introgressed Neanderthal haplotypes (Chen et al. 2020), so-called deserts of introgression, distributed across only four chromosomes (partially overlapping previous independent research, i.e. Vernot et al. 2016).

9780192886491_Book.indb 215 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:44

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Intriguingly, one desert corresponds to a 17Mbp region in chromosome 7, likely under strong purifying selection in *Homo sapiens* but not in other populations (Kuhlwilm 2018), and where *FOXP2* is found, hinting to possible differences in brain function. However, it is a large region and encompasses several other genes, making it necessary to further investigate the functional underpinnings of genetic differences within this region. A special case of depletion of introgression is the X chromosome (Sankararaman et al. 2014; also seen in other species; Pawar et al. 2023), which contains almost an order of magnitude fewer introgressed fragments from both Neanderthals and Denisovans than the autosomes (Chevy et al. 2023). It is still unclear whether or to which degree this is driven by sex-biased migration patterns, hybrid incompatibilities, and selection either against archaic variants or in favour of later acquired variants (Skov et al. 2023). However, it is possible that genetic variants on the X chromosome contribute to human-specific features related to reproduction or other traits. In conclusion, introgression deserts generally, and on this chromosome in particular, are still poorly understood.

The selected examples above show that a range of evidence for the impact of *Homo* species-specific variants point to phenotypic divergences taking place at early stages of neurodevelopment. High-throughput sequencing technologies and experimental models unravel key aspects of human brain development in high spatiotemporal resolution (see section 12.3). This yields a powerful approach to begin to formulate explanatory hypotheses linking species-specific genotypes and phenotypic differences. In this context, it is worth mentioning the evidence supporting a species-specific brain ontogenetic trajectory linked to one of the most distinctive anatomical traits of *Homo sapiens*, absent in our closest relatives: a globular neurocranium. Based on three-dimensional reconstructions of the endocranial shape of *Homo sapiens*, chimpanzees, and Neanderthals from computed tomography encompassing different developmental windows (from birth to adulthood), comparative morphometric analysis has provided evidence for a globularization phase at perinatal stages taking place exclusively in the *Homo sapiens* lineage, possibly linked to an expansion of the cerebellar and parietal regions (Neubauer et al. 2010; Gunz et al. 2010, 2012). Additionally, evidence from *Homo sapiens* fossils at different evolutionary time points suggest that this globular trait emerged gradually and after 100 thousand years ago (kya), thus being closely connected to the emergence and expansion of anatomically modern humans (Neubauer et al. 2018).

Prominently, in a study based on neuroimaging genomic data, Gunz et al. (2019) elaborated a measure of endocranial globularity from MRI scans of thousands of present-day individuals (mostly from Europe) and found that introgressed Neanderthal variants in two genomic regions were statistically associated to reduced endocranial globularity. Some of these variants were predicted to impact gene expression regulation, and more significantly the expression of *UBR4* in the putamen and *PHLPP1* in the cerebellum, perhaps influencing neurogenesis and myelination, respectively (Gunz et al. 2019). This line of research shifts the focus from size (where Neanderthals fall within or above present-day human variation) to shape, as well as encouraging us to consider other regions beyond the neocortex. One such region is the cerebellum, whose derived status in modern humans acquires special prominence in the context of the evolution

9780192886491_Book.indb 216 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:44

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

of the faculty of language, as the dominant cerebellar hemisphere for language has been found significantly enlarged in *Homo sapiens* when compared to the Neanderthals (Kochiyama et al. 2018). Hypotheses on the emergence of modern human cognition can be now traced back to genetic mutations revealed by palaeogenomic analysis and then mapped into specific developmental processes implicated in brain growth. As such, advances in modelling of the early stages of brain development *in vitro* are critical to make these hypotheses experimentally tractable.

12.3 Experimental interrogation of the functional relevance of *Homo*-specific genetic variants

While no primary brain tissue will ever be available for our extinct relatives, there are also ethical limitations to the availability of primary tissue for analysis as well as experimentation. This restricted accessibility to human brain tissue for experimentation stands as a significant obstacle for the elucidation of developmental and evolutionary aspects of our species' brain, especially those that might not be faithfully recapitulated with the use of animal models. In recent years, new technologies have been developed to model aspects of nervous system development and function. Prominently among these, brain organoid models stand out as powerful systems to mimic the intricate three-dimensional arrangement and cellular and molecular complexities of dynamic tissues such as the developing human brain, generally with the use of induced pluripotent stem cells (iPSCs), including from non-human primate species (Velasco et al. 2020). Numerous protocols are currently available and being developed to produce unguided neural organoids, or guided organoids to recapitulate developmental processes from specific structures such as the cerebral cortex or the spinal cord (Paşca et al. 2022). Moreover, experimental techniques applicable to *in vitro* models for the functional investigation of evolutionary-relevant mutations have been developed, such as massively parallel reporter assays (MPRAs) or the targeted modification of single nucleotides in cell lines using CRISPR-Cas systems (Pollen et al. 2023). The magnitude of non-synonymous mutations that distinguish our species from the Neanderthals and Denisovans makes their functional interrogation experimentally tractable, and the elucidation of the functional roles of the larger number of regulatory variants has already started (Weiss et al. 2021; Jagoda et al. 2022; Rong et al. 2023).

Datasets on variation and functional data are growing, and machine learning-based approaches are increasingly considered in genomics (Whalen et al. 2023; Huang et al. 2024). The development of new computational tools for data analysis and particularly prediction is fundamental for our understanding of the evolutionary consequences of species-specific mutations (Whalen et al. 2023), and key for the formulation of testable hypotheses which could be tackled using the abovementioned experimental

9780192886491_Book.indb 217 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:44

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

approaches. In fact, it is now possible to comparatively evaluate sequence divergence informed by palaeogenomics and reconstruct past regulatory scenarios such as alternative splicing or chromatin folding (McArthur et al. 2022; Brand et al. 2023). Prominently, patterns of cytosine decay in ancient DNA were used to infer methylation maps from the genome sequences of Neanderthals and Denisovans (Gokhman et al. 2014). These efforts led to a tentatively outlined anatomical traits of Denisovans, and uncovered differentially methylated regions likely impacting gene expression levels and linked to *Homo* lineage-specific differences (Gokhman et al. 2019, 2020). Indeed, differentially methylated regions derived in *Homo sapiens* were predicted to specifically affect facial regions, larynx, and pelvis (Gokhman et al. 2020).

12.3.1 Molecular and cellular substrates of species-specific brain ontogenetic trajectories

In the quest to uncover relevant molecular and cellular processes that might have contributed to brain ontogenetic divergences before the split of *Homo sapiens* and Neanderthal/Denisovan lineages, significant efforts have been devoted to identify novel genes specific to the human lineage, regulatory variants impacting gene regulatory networks or differences in progenitor cell populations (see, for detailed reviews, Libé-Philippot and Vanderhaeghen 2021; Pinson et al. 2022; Vanderhaeghen and Polleux 2023). Remarkable progress was facilitated by the introduction of high-throughput RNA sequencing technologies (Romero et al. 2012), further propelled by multidisciplinary approaches that enable us to comprehensively characterize region- and cell-type specific features of human brain development and function (consider, for instance, the BRAIN Initiative Cell Census Network; Winnubst and Arber 2021). Recent comparative genomics studies based on both brain primary tissue and organoid models across primate species have provided insights into divergent gene expression profiles at different stages of neurodevelopment (Pollen et al. 2019; Kanton et al. 2019), and have revealed that structures outside the neocortex such as the cerebellum or hypothalamus, and cell types such as astrocytes and oligodendrocyte progenitors, report rapid evolution in the human lineage (Khrameeva et al. 2020). Additionally, a recent appreciation of developmental timing as key mechanism explaining species differences in neural progenitor cell behaviour has emerged (Mora-Bermúdez et al. 2016; Otani et al. 2016; Iwata et al. 2023). For instance, a prolonged prometaphase-metaphase in apical progenitors in humans in comparison to other great apes is associated to an increased proliferative capacity (Mora-Bermúdez et al. 2016). Another developmental process possibly linked to the dramatic increase in neocortical surface in the *Homo* lineage pertains to a delayed transition between two progenitor cell types at early neurodevelopmental stages: comparative brain organoid modelling across primate species revealed a delayed transition of neuroepithelial cells to radial glia cells (two types of progenitor cells found in the germinal zones during brain development) in humans, predicting a 1.9-fold increase in the number of progenitors and neurons in comparison to other apes

9780192886491_Book.indb 218 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:44

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

(Benito-Kwiecinski et al. 2021). A zinc-finger transcription factor, ZEB2, was reported as a plausible critical regulator of this transition, its expression peaking five days later in human derived organoids than in gorillas (Benito-Kwiecinski et al. 2021). Intriguingly, the gene encoding for ZEB2 has been linked to human accelerated regions, raising the question whether human accelerated regions contribute to its differential regulation finally impacting cerebral cortex expansion (Benito-Kwiecinski et al. 2021).

How has palaeogenomic research contributed to the experimental interrogation of the genetic basis of traits derived in *Homo sapiens*? First, inquiries into the genomes from our closest extinct relatives revealed that a substantial proportion of derived missense variants in our species are associated to genes implicated in the neural progenitor cell division (Prüfer et al. 2014). One of these missense variants corresponds to a lysine-to-arginine substitution that distinguishes *Homo sapiens* and Neanderthals *TKTL1* gene (a variant however present at around 32 per cent frequency in the Khoe-San southern Africa group; Zeberg et al. 2024). Using a variety of experimental approaches in human primary tissue, mice, ferret, and brain organoids, Pinson et al. (2022) mechanistically linked the *TKTL1* non-synonymous mutation to metabolic pathways involved in fatty acid synthesis, selectively impacting a specific neural progenitor cell type (outer radial glial cells), amplifying its proliferative capabilities with a concomitant increase in cortical neurogenesis. While Pinson et al. (2022) further constrained the impact of such mutation to the frontal lobe, how to interpret these results in light of the known phenotypic differences in brain organization, as inferred from the fossil record, remains unclear. Another metabolic pathway impacted in recent human evolution pertains to a single nucleotide variant derived in the *Homo sapiens* lineage causing an alanine-to-valine substitution in the adenylosuccinate lyase enzyme ADSL (Stepanova et al. 2021). Comparatively to the Neanderthals protein version, the modern human ADSL protein is less stable and results in a decreased synthesis of purines, more prominently in the brain (Stepanova et al. 2021).

Another set of genes brought by palaeogenomic studies are those related to the spindle apparatus and the kinetochore. In fact, spindle-associated genes, required for the correct placement and segregation of chromosomes, accumulated an excess of missense mutations in our lineage (Peyrégne et al. 2022). Contrasting the functional effects of modern human and Neanderthal mutations on three spindle-associated genes in mice and brain organoid models, Mora-Bermúdez et al. (2022) discovered that the modern human-specific allele versions in two genes—*KIF18A* and *KNL1*—cause a longer metaphase with few chromosome segregation errors in neural progenitors when compared to the Neanderthal alleles. Similar experimental manipulations on a third gene, *SPAG5* revealed no significant functional impact due to the derived allele in *Homo sapiens*, although it was detected in a genomic region that shows signals of positive selection in modern humans (Peyrégne et al. 2022; Pinson et al. 2022). The genomic divergence observed in spindle genes during recent human evolution also provides more lessons on our shared history with Neanderthals. Substitutions in one gene, the kinetochore scaffold 1 gene *KNL1*, previously thought to be unique to the modern human lineage, were later found to be also present in some Neanderthal individuals, likely as

9780192886491_Book.indb 219 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:44

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

the result of interbreeding with ancestral modern human populations after 265 kya (Kuhlwilm et al., 2016; Peyrégne et al. 2022). Finally, the regulator of alternative splicing *NOVA1* carries an ancestral missense mutation in the Neanderthals and Denisovans shown to slow the pace of cortical organoid development and to impact synaptogenesis, suggesting possible differences in neural network connectivity among *Homo* species (Trujillo et al. 2021; see also Maricic et al. 2021, Herai et al. 2021).

12.3.2 A study case: The self-domestication hypothesis

While, in this chapter, we have devoted our attention to the evolutionary differences that contribute to explaining the acquisition of the modern human brain, other developmental mechanisms and systems might have been modified over the course of *Homo sapiens*' evolution. The self-domestication hypothesis refers to the idea that certain aspects of the behavioural profile of our species, such as a marked reduction in reactive aggression and a distinctively prosocial phenotype, are linked to phenotypic traits reminiscent to that of domesticated species when compared to their wild counterparts. The traits that encompass the so-called domestication syndrome—to a varied degree and asymmetrical distribution across species—include reduced body and brain size, attenuation of sex differences, flattened and smaller faces, or depigmentation (Wilkins et al. 2014). The self-domestication hypothesis states that the evolutionary trajectory leading to our species' domestication enabled the emergence of evolved traits characteristic of our species, and perhaps most significantly, of language (Okanoya 2017; Thomas and Kirby 2018; O'Rourke et al. 2021; Wrangham 2021; Boeckx 2023). Progress in the genotype–phenotype mapping problem was achieved with the proposition that the domestication syndrome arises from a diminished production and migration of neural crest cells (cell types responsible for the generation of disparate types of tissues, such as the facial bones or the adrenal glands; Wilkins et al. 2014). This provided a mechanistic explanation whose predictions are experimentally tractable. The self-domestication hypothesis has equally benefitted from the increasing number of high-coverage genomes from extant and extinct species, which now afford the interrogation of the molecular underpinnings of the hypothesized mild neurocristopathy affecting domesticated animals comparatively across a wide range of species. In recent work, Zanella et al. (2019) put forward an experimental design to test whether the transcriptional regulatory networks involved in neural crest cells proliferation and differentiation, and affected in human neurocristopathies, were under selection in anatomically modern humans. Zanella et al. (2019) made use of a large cohort of iPSC lines derived from patients clinically diagnosed with Williams-Beuren syndrome and Williams-Beuren region duplication syndrome, as a result of copy number variations at 7q11.23. This allowed the molecular and cellular dissection of the consequences of BAZ1B dosage imbalance, a master regulator of neural crest cells, from patients whose phenotypic consequences bear direct resemblance to key traits in anatomically modern humans: a retracted face and prosocial behaviour. The palaeogenomic inquiry of the regulatory

9780192886491_Book.indb 220 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:44

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

circuits underlying neural cress cell functioning in these models revealed that the BAZ1B regulatory network linked to neurocristopathic facial dysmorphisms underwent evolutionary modifications specific to anatomically modern humans, providing, in this way, evidence for the molecular correlates behind the craniofacial traits at the core of the self-domestication hypothesis (Zanella et al. 2019). This approach exemplifies how neurodisease modelling in combination with palaeogenomic analyses can be used to experimentally test longstanding hypotheses on the evolution of modern humans, especially when combined with comparative genomics research (in this case, domesticated species vs. their wild counterparts, which also exploits progress in ancient genomics).

12.4 Conclusion

After more than a decade of the first draft of a Neanderthal genome, palaeogenomic research has greatly advanced our understanding of the origin and essence of our species, complementing what can be investigated on the basis of comparative research using the genomes of a wide range of living species. This research contributes to necessarily interdisciplinary approaches to unravel the genetic basis of the so-called human condition. Assumptions of a single region of origin for our species within Africa are unlikely to be correct, where evidence from the fossil record favours a more complex, mosaic-like pattern for the emergence of key anatomical characteristics of our species (Scerri et al. 2018). Likewise, the rise of behavioural modernity might not be considered as a monolithic entity emerging at a specific time point and geographic area, but rather the result of a multifactorial gradual process (Scerri and Will 2023). Promising technological advances, such as the analysis of ancient DNA from cave sediments (Slon et al. 2017), will continue to transform our ability to resurrect past scenarios of human evolution, along with a more diverse representation of the archaeological record. Results derived from this line of research will be best complemented with insights from comparative genomics, as has been done in the case of vocal learning (Fisher 2019).

References

Alagöz, G., Molz, B., Eising, E., et al. (2022). 'Using neuroimaging genomics to investigate the evolution of human brain structure', *Proceedings of the National Academy of Sciences* 119: e2200638119.

Andirkó, A., Moriano, J., Vitriolo, A., et al. (2022). 'Temporal mapping of derived high-frequency gene variants supports the mosaic nature of the evolution of Homo sapiens', *Scientific Reports* 12: 9937.

Atkinson, E. G., Audesse, A. J., Palacios, J. A., et al. (2018). 'No evidence for recent selection at FOXP2 among diverse human populations', *Cell* 174: 1424–1435.e15.

Auton, A., Abecasis, G. R., Altshuler, D. M., et al. (2015). 'A global reference for human genetic variation', *Nature* 526: 68–74.

9780192886491_Book.indb 221 /12_first_proofs/first_pr[illegible]/xml_for_typesetting 27-Mar-25 23:02:44

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Benito-Kwiecinski, S., Giandomenico, S. L., Sutcliffe, M., et al. (2021). 'An early cell shape transition drives evolutionary expansion of the human forebrain', *Cell* 184: 2084–2102.e19.

Bergström, A., Stringer, C., Hajdinjak, M., et al. (2021). 'Origins of modern human ancestry', *Nature* 590: 229–237.

Besenbacher, S., Hvilsom, C., Marques-Bonet, T., et al. (2019). 'Direct estimation of mutations in great apes reconciles phylogenetic dating', *Nature Ecology & Evolution* 3: 286–292.

Boeckx, C. (2023). 'What made us "hunter-gatherers of words"', *Frontiers in Neuroscience* 17: 1080861.

Brand, C. M., Colbran, L. L., and Capra, J. A. (2023). 'Resurrecting the alternative splicing landscape of archaic hominins using machine learning', *Nature Ecology & Evolution* 7: 939–953.

Briscoe, J. and Marín, O. (2020). 'Looking at neurodevelopment through a big data lens', *Science (New York, N.Y.)* 369 eaaz8627.

Buisan, R., Moriano, J., Andirkó, A., and Boeckx, C. (2022). 'A brain region-specific expression profile for genes within large introgression deserts and under positive selection in Homo sapiens', *Frontiers in Cell and Developmental Biology* 10: 824740.

Chabout, J., Sarkar, A., Patel, S. R., et al. (2016). 'A Foxp2 mutation implicated in human speech deficits alters sequencing of ultrasonic vocalizations in adult male mice', *Frontiers in Behavioral Neuroscience* 10: 197.

Chen, L., Wolf, A. B., Fu, W., et al. (2020). 'Identifying and interpreting apparent Neanderthal ancestry in African individuals', *Cell* 180: 677–687.e16.

Chevy, E. T., Huerta-Sánchez, E., and Ramachandran, S. (2023). 'Integrating sex-bias into studies of archaic introgression on chromosome X', *PLOS Genetics* 19: e1010399.

Coppo, L., Mishra, P., Siefert, N., et al. (2022). 'A substitution in the glutathione reductase lowers electron leakage and inflammation in modern humans', *Science Advances* 8(1): eabm1148.

de Boer, B., Thompson, B., Ravignani, A., and Boeckx, C. (2020). 'Evolutionary dynamics do not motivate a single-mutant theory of human language', *Scientific Reports* 10: 451.

Enard, W., Przeworski, M., Fisher, S. E., et al. (2002). 'Molecular evolution of FOXP2, a gene involved in speech and language', *Nature* 418: 869–872.

Fisher, S. E. (2019). 'Human genetics: The evolving story of FOXP2', *Current Biology* 29: R65–R67.

Fontsere, C., de Manuel, M., Marques-Bonet, T., and Kuhlwilm, M. (2019). 'Admixture in mammals and how to understand its functional implications', *BioEssays* 41(12): 1900123.

Gokhman, D., Lavi, E., Prüfer, K., et al. (2014). 'Reconstructing the DNA methylation maps of the Neandertal and the Denisovan', *Science (New York, N.Y.)* 344: 523–527.

Gokhman, D., Mishol, N., de Manuel, M., et al. (2019). 'Reconstructing Denisovan anatomy using DNA methylation maps', *Cell* 179: 180–192.e10.

Gokhman, D., Nissim-Rafinia, M., Agranat-Tamir, L., et al. (2020). 'Differential DNA methylation of vocal and facial anatomy genes in modern humans', *Nature Communications* 11: 1189.

Green, R. E., Krause, J., Ptak, S. E., et al. (2006). 'Analysis of one million base pairs of Neanderthal DNA', *Nature* 444: 330–336.

Green, R. E., Krause, J., Briggs, A. W., et al. (2010). 'A draft sequence of the Neandertal genome', *Science (New York, N.Y.)* 328: 710–722.

Gunz, P., Neubauer, S., Maureille, B., and Hublin, J.-J. (2010). 'Brain development after birth differs between Neanderthals and modern humans', *Current Biology* 20: R921–R922.

Gunz, P., Neubauer, S., Golovanova, L., et al. (2012). 'A uniquely modern human pattern of endocranial development: Insights from a new cranial reconstruction of the Neandertal newborn from Mezmaiskaya', *Journal of Human Evolution* 62: 300–313.

9780192886491_Book.indb 222 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:44

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Gunz, P., Tilot, A. K., Wittfeld, K., et al. (2019). 'Neandertal introgression sheds light on modern human endocranial globularity', *Current Biology* 29: 120–127.e5.

Hajdinjak, M., Fu, Q., Hübner, A., et al. (2018). 'Reconstructing the genetic history of late Neanderthals', *Nature* 555: 652–656.

Hajdinjak, M., Mafessoni, F., Skov, L., et al. (2021). 'Initial Upper Palaeolithic humans in Europe had recent Neanderthal ancestry', *Nature* 592: 253–257.

Herai, R. H., Szeto, R. A., Trujillo, C. A., and Muotri, A. R. (2021). 'Response to comment on "Reintroduction of the archaic variant of NOVA1 in cortical organoids alters neurodevelopment"', *Science* 374: eabi9881.

Huang, X., Rymbekova, A., Dolgova, O., et al. (2024). 'Harnessing deep learning for population genetic inference', *Nature Reviews Genetics* 25: 61–78.

Iwata, R., Casimir, P., Erkol, E., et al. (2023). 'Mitochondria metabolism sets the species-specific tempo of neuronal development', *Science* 379: eabn4705.

Jagoda, E., Xue, J. R., Reilly, S. K., et al. (2022). 'Detection of Neanderthal adaptively introgressed genetic variants that modulate reporter gene expression in human immune cells', *Molecular Biology and Evolution* 39: msab304.

Kanton, S., Boyle, M. J., He, Z., et al. (2019). 'Organoid single-cell genomic atlas uncovers human-specific features of brain development,' *Nature* 574: 418–422.

Khrameeva, E., Kurochkin, I., Han, D., et al. (2020). 'Single-cell-resolution transcriptome map of human, chimpanzee, bonobo, and macaque brains', *Genome Research* 30: 776–789.

King, M. C. and Wilson, A. C. (1975). 'Evolution at two levels in humans and chimpanzees', *Science (New York, N.Y.)* 188: 107–116.

Kochiyama, T., Ogihara, N., Tanabe, H. C., et al. (2018). 'Reconstructing the Neanderthal brain using computational anatomy', *Scientific Reports* 8: 6296.

Krause, J., Lalueza-Fox, C., Orlando, L., et al. (2007). 'The derived FOXP2 variant of modern humans was shared with Neandertals', *Current Biology* 17: 1908–1912.

Krings, M., Stone, A., Schmitz, R. W., et al. (1997). 'Neandertal DNA sequences and the origin of modern humans', *Cell* 90: 19–30.

Kuderna, L. F. K., Gao, H., Janiak, M. C., et al. (2023). 'A global catalog of whole-genome diversity from 233 primate species', *Science* 380: 906–913.

Kuhlwilm, M. (2018). 'The evolution of FOXP2 in the light of admixture', *Current Opinion in Behavioral Sciences* 21: 120–126.

Kuhlwilm, M. and Boeckx, C. (2019). 'A catalog of single nucleotide changes distinguishing modern humans from archaic hominins', *Scientific Reports* 9: 8463.

Kuhlwilm, M., Gronau, I., Hubisz, M. J., de Filippo, C., Prado-Martinez, J., Kircher, M., Fu, Q., Burbano, H. A., Lalueza-Fox, C., de la Rasilla, M., Rosas, A., Rudan, P., Brajkovic, D., Kucan, Ž., Gušic, I., Marques-Bonet, T., Andrés, A. M., Viola, B., Pääbo, S., Meyer, M., … Castellano, S. (2016). 'Ancient gene flow from early modern humans into Eastern Neanderthals'. *Nature* 530(7591): 429–433, https://doi.org/10.1038/nature16544.

Lai, C. S. L., Fisher, S. E., Hurst, J. A., et al. (2001). 'A forkhead-domain gene is mutated in a severe speech and language disorder', *Nature* 413: 519–523.

Libé-Philippot, B. and Vanderhaeghen, P. (2021). 'Cellular and molecular mechanisms linking human cortical development and evolution', *Annual Review of Genetics* 55(1): 555–581, eprint: https://doi.org/10.1146/annurev-genet-071719-020705.

Liégeois, F., Baldeweg, T., Connelly, A., et al. (2003). 'Language fMRI abnormalities associated with FOXP2 gene mutation', *Nature Neuroscience* 6: 1230–1237.

Mafessoni, F., Grote, S., de Filippo, C., et al. (2020). 'A high-coverage Neandertal genome from Chagyrskaya Cave', *Proceedings of the National Academy of Sciences* 117: 15132–15136.

9780192886491_Book.indb 223 /12_first_proofs/first_pr…/xml_for_typesetting 27-Mar-25 23:02:44

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Mallick, S., Li, H., Lipson, M., et al. (2016). 'The Simons Genome Diversity Project: 300 genomes from 142 diverse populations', *Nature* 538: 201–206.

Maricic, T., Günther, V., Georgiev, O., et al. (2013). 'A recent evolutionary change affects a regulatory element in the human FOXP2 gene', *Molecular Biology and Evolution* 30: 844–852.

Maricic, T., Helmbrecht, N., Riesenberg, S., et al. (2021). 'Comment on "Reintroduction of the archaic variant of NOVA1 in cortical organoids alters neurodevelopment"', *Science* 374: eabi6060.

McArthur, E., Rinker, D. C., Gilbertson, E. N., et al. (2022). 'Reconstructing the 3D genome organization of Neanderthals reveals that chromatin folding shaped phenotypic and sequence divergence', Pages: 2022.02.07.479462 Section: New Results.

McCoy, R. C., Wakefield, J., and Akey, J. M. (2017). 'Impacts of Neanderthal-introgressed sequences on the landscape of human gene expression', *Cell* 168: 916–927.e12.

Meyer, M., Kircher, M., Gansauge, M.-T., et al. (2012). 'A high-coverage genome sequence from an archaic Denisovan individual', *Science (New York, N.Y.)* 338: 222–226.

Mora-Bermúdez, F., Badsha, F., Kanton, S., et al. (2016). 'Differences and similarities between human and chimpanzee neural progenitors during cerebral cortex development', *eLife* 5: e18683.

Mora-Bermúdez, F., Kanis, P., Macak, D., et al. (2022). 'Longer metaphase and fewer chromosome segregation errors in modern human than Neanderthal brain development', *Science Advances* 8: eabn7702.

Murugan, M., Harward, S., Scharff, C., and Mooney, R. (2013). 'Diminished FoxP2 levels affect dopaminergic modulation of corticostriatal signaling important to song variability', *Neuron* 80: 1464–1476.

Neubauer, S., Gunz, P., and Hublin, J.-J. (2010). 'Endocranial shape changes during growth in chimpanzees and humans: A morphometric analysis of unique and shared aspects', *Journal of Human Evolution* 59: 555–566.

Neubauer, S., Hublin, J.-J., and Gunz, P. (2018). 'The evolution of modern human brain shape', *Science Advances* 4: eaao5961.

Okanoya, K. (2017). 'Sexual communication and domestication may give rise to the signal complexity necessary for the emergence of language: An indication from songbird studies', *Psychonomic Bulletin & Review* 24: 106–110.

Orlando, L., Allaby, R., Skoglund, P., et al. (2021). 'Ancient DNA analysis', *Nature Reviews Methods Primers* 1: 1–26.

O'Rourke, T., Martins, P. T., Asano, R., et al. (2021). 'Capturing the effects of domestication on vocal learning complexity', *Trends in Cognitive Sciences* 25: 462–474.

Otani, T., Marchetto, M. C., Gage, F. H., et al. (2016). '2D and 3D stem cell models of primate cortical development identify species-specific differences in progenitor behavior contributing to brain size', *Cell Stem Cell* 18: 467–480.

Pääbo, S. (2014). 'The human condition—A molecular approach', *Cell* 157: 216–226.

Pääbo, S., Higuchi, R. G., and Wilson, A. C. (1989). 'Ancient DNA and the polymerase chain reaction: The emerging field of molecular archaeology', *The Journal of Biological Chemistry* 264: 9709–9712.

Pașca, S. P., Arlotta, P., Bateup, H. S., et al. (2022). 'A nomenclature consensus for nervous system organoids and assembloids', *Nature* 609: 907–910.

Pawar, H., Rymbekova, A., Cuadros-Espinoza, S., et al. (2023). 'Ghost admixture in eastern gorillas', *Nature Ecology & Evolution* 7: 1503–1514.

9780192886491_Book.indb 224 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:45

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Petr, M., Pääbo, S., Kelso, J., and Vernot, B. (2019). 'Limits of long-term selection against Neandertal introgression', *Proceedings of the National Academy of Sciences* 116: 1639–1644.
Peyrégne, S., Boyle, M. J., Dannemann, M., and Prüfer, K. (2017). 'Detecting ancient positive selection in humans using extended lineage sorting', *Genome Research* 27: 1563–1572.
Peyrégne, S., Kelso, J., Peter, B. M., and Pääbo, S. (2022). 'The evolutionary history of human spindle genes includes back-and-forth gene flow with Neandertals', *eLife* 11: e75464.
Pinson, A., Xing, L., Namba, T., et al. (2022). 'Human TKTL1 implies greater neurogenesis in frontal neocortex of modern humans than Neanderthals', *Science* 377: eabl6422.
Poeppel, D. (2012). 'The maps problem and the mapping problem: Two challenges for a cognitive neuroscience of speech and language', *Cognitive Neuropsychology* 29(1–2): 34–55.
Pollen, A., Bhaduri, A., Andrews, M. G., et al. (2019). 'Establishing cerebral organoids as models of human-specific brain evolution', *Cell* 176: 743–756.e17.
Pollen, A., Kilik, U., Lowe, C. B., and Camp, J. G. (2023). 'Human-specific genetics: New tools to explore the molecular and cellular basis of human evolution', *Nature Reviews Genetics* 1–25.
Prüfer, K., Racimo, F., Patterson, N., et al. (2014). 'The complete genome sequence of a Neanderthal from the Altai Mountains', *Nature* 505: 43–49.
Prüfer, K., de Filippo, C., Grote, S., et al. (2017). 'A high-coverage Neandertal genome from Vindija Cave in Croatia', *Science (New York, N.Y.)* 35: 655–658.
Racimo, F. (2016). 'Testing for ancient selection using cross-population allele frequency differentiation', *Genetics* 202: 733–750.
Reich, D., Green, R. E., Kircher, M., et al. (2010). 'Genetic history of an archaic hominin group from Denisova Cave in Siberia', *Nature* 468: 1053–1060.
Reilly, P. F., Tjahjadi, A., Miller, S. L., et al. (2022). 'The contribution of Neanderthal introgression to modern human traits', *Current Biology* 32: R970–R983.
Romero, G., Ruvinsky, I., and Gilad, Y. (2012). 'Comparative studies of gene expression and the evolution of gene regulation', *Nature Reviews Genetics* 13: 505–516.
Rong, S., Neil, C. R., Welch, A., et al. (2023). 'Large-scale functional screen identifies genetic variants with splicing effects in modern and archaic humans', *Proceedings of the National Academy of Sciences* 120: e2218308120.
Sadakata, T., Washida, M., Iwayama, Y., et al. (2007). 'Autistic-like phenotypes in Cadps2-knockout mice and aberrant CADPS2 splicing in autistic patients', *The Journal of Clinical Investigation* 117: 931–943.
Sadakata, T., Shinoda, Y., Ishizaki, Y., and Furuichi, T. (2017). 'Analysis of gene expression in Ca2+-dependent activator protein for secretion 2 (Cadps2) knockout cerebellum using GeneChip and KEGG pathways', *Neuroscience Letters* 639: 88–93.
Sankararaman, S., Mallick, S., Dannemann, M., et al. (2014). 'The genomic landscape of Neanderthal ancestry in present-day humans', *Nature* 507: 354–357.
Scerri, E. M. L. and Will, M. (2023). 'The revolution that still isn't: The origins of behavioral complexity in Homo sapiens', *Journal of Human Evolution* 179: 103358.
Scerri, E. M. L., Thomas, M. G., Manica, A., et al. (2018). 'Did our species evolve in subdivided populations across Africa, and why does it matter?', *Trends in Ecology & Evolution* 33: 582–594.
Skov, L., Peyrégne, S., Popli, D., et al. (2022). 'Genetic insights into the social organization of Neanderthals', *Nature* 610: 519–525.
Skov, L., Coll Macià, M., Lucotte, E. A., et al. (2023). 'Extraordinary selection on the human X chromosome associated with archaic admixture', *Cell Genomics* 3: 100274.
Slon, V., Hopfe, C., Weiß, C. L., et al. (2017). 'Neandertal and Denisovan DNA from Pleistocene sediments', *Science* 356: 605–608.

9780192886491_Book.indb 225 /12_first_proofs/first_pr[illegible]/xml_for_typesetting 27-Mar-25 23:02:45

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Slon, V., Viola, B., Renaud, G., et al. (2017). 'A fourth Denisovan individual', *Science Advances* 3: e1700186.

Stepanova, V., Moczulska, K. E., Vacano, G. N., et al. (2021). 'Reduced purine biosynthesis in humans after their divergence from Neandertals', *eLife* 10: e58741.

Thomas, J. and Kirby, S. (2018). 'Self domestication and the evolution of language', *Biology & Philosophy* 33: 9.

Trujillo, C. A., Rice, E. S., Schaefer, N. K., et al. (2021). 'Reintroduction of the archaic variant of NOVA1 in cortical organoids alters neurodevelopment', *Science* 371: eaax2537.

Vanderhaeghen, P. and Polleux, F. (2023). 'Developmental mechanisms underlying the evolution of human cortical circuits', *Nature Reviews Neuroscience* 24: 213–232.

Varki, A. and Altheide, T. K. (2005). 'Comparing the human and chimpanzee genomes: Searching for needles in a haystack', *Genome Research* 15: 1746–1758.

Velasco, S., Paulsen, B., and Arlotta, P. (2020). '3D brain organoids: Studying brain development and disease outside the embryo', *Annual Review of Neuroscience* 43(1): 375–389, eprint: https://doi.org/10.1146/annurev-neuro-070918-050154.

Veller, C., Edelman, N. B., Muralidhar, P., and Nowak, M. A. (2023). 'Recombination and selection against introgressed DNA', *Evolution* 77: 1131–1144.

Vernot, B., Tucci, S., Kelso, J., et al. (2016). 'Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals', *Science (New York, N.Y.)* 352: 235–239.

Wall, J. D. and Kim, S. K. (2007). 'Inconsistencies in Neanderthal genomic DNA sequences', *PLOS Genetics* 3: e175.

Watkins, K. E., Vargha-Khadem, F., Ashburner, J., et al. (2002). 'MRI analysis of an inherited speech and language disorder: Structural brain abnormalities', *Brain* 125: 465–478.

Weiss, C. V., Harshman, L., Inoue, F., et al. (2021). 'The cis-regulatory effects of modern human-specific variants', *eLife* 10: e63713.

Whalen, S., Inoue, F., Ryu, H., et al. (2023). 'Machine learning dissection of human accelerated regions in primate neurodevelopment', *Neuron* 11: 857–873.e8.

Wilkins, S., Wrangham, R. W., and Fitch, W. T. (2014). 'The "domestication syndrome" in mammals: A unified explanation based on neural crest cell behavior and genetics', *Genetics* 197: 795–808.

Winnubst, J. and Arber, S. (2021). 'A census of cell types in the brain's motor cortex', *Nature* 598: 33–34.

Wrangham, R. W. (2021). 'Targeted conspiratorial killing, human self-domestication and the evolution of groupishness', *Evolutionary Human Sciences* 3: e26.

Zanella, M., Vitriolo, A., Andirko, A., et al. (2019). 'Dosage analysis of the 7q11.23 Williams region identifies BAZ1B as a major human gene patterning the modern human face and underlying self-domestication', *Science Advances* 5: eaaw7908.

Zeberg, H., Dannemann, M., Sahlholm, K., et al. (2020a). 'A Neanderthal sodium channel increases pain sensitivity in present-day humans', *Current Biology: CB* 30: 3465–3469.e.

Zeberg, H., Kelso, J., and Pääbo, S. (2020b). 'The Neandertal progesterone receptor', *Molecular Biology and Evolution* 37: 2655–2660.

Zeberg, H., Jakobsson, M., and Pääbo, S. (2024). 'The genetic changes that shaped Neandertals, Denisovans, and modern humans', *Cell* S0092–8674(23)01403–4.

Zhang, J., Webb, D. M., and Podlaha, O. (2002). 'Accelerated protein evolution and origins of human-specific features: FOXP2 as an example', *Genetics* 162: 1825–1835.

9780192886491_Book.indb 226 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:45

CHAPTER 13

JOINT ANALYSIS OF HUMAN LINGUISTIC AND GENOMIC VARIATION

YAKOV PICHKAR AND NICOLE CREANZA

13.1 Introduction

The history of languages includes both a history of ideas and of the people who have them. Both language and genetics are heritable forms of human variation, and they contain parallel histories of populations, their movements, and their interactions (Darwin 1871; Cavalli-Sforza and Feldman 1981; Lumsden and Wilson 1981; Boyd and Richerson 1985). Neither language nor genetics contains a complete record of this ever-unfolding process, but their joint analysis can clarify the web of historical relations. This type of analysis presents numerous challenges, since genetic variation arises from biologically inherited genes, whereas languages are composed of socially transmitted ideas (Guglielmino et al. 1995). However, to evaluate evidence for various hypotheses describing the history of peoples, several methods have been applied to study the relationship between genetics and language (Cavalli-Sforza et al. 1992; Haber et al. 2013; Creanza et al. 2015; Di Corcia et al. 2017). In this chapter, we summarize existing literature comparing genetic and linguistic variation, with an emphasis on a particular method: Procrustes analysis.

The synthesis of biological and linguistic history began with Darwin's observation that language and its features, much like biological traits, are heritable, leading him to hypothesize that knowledge of linguistic history could disentangle the histories connecting groups of people (Darwin 1871). In *On the Origin of Species*, he stated that 'a perfect pedigree of mankind… would afford the best classification of the various languages now spoken throughout the world'. Since then, comparisons between population-level genetic variation and various aspects of human languages have become not only possible but detailed and quantitative. These analyses of human linguistic

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

and genetic variation have shown not only linguistic parallels to the 'pedigree of mankind'—i.e. human genetic phylogenies—but also linguistic signatures of the complex interactions between more distantly related populations.

As the number of genetic sequences has grown, joint genetic-linguistic analysis can improve our understanding of the interactions that occurred between groups of people (Barbujani and Sokal 1990, 1991). At the scale of individuals, genetic and linguistic variation are generated and distributed dissimilarly: genes are inherited from parents to children in a process of vertical transmission, whereas linguistic features can arise and spread within one generation, and any given individual might learn a language through many interactions with a multitude of people (Guglielmino et al. 1995). At the larger scale of communities and populations, group dynamics can be structured by homophily—people preferentially interacting with those similar to themselves—and by geography, both of which can lead to a correlation between genes and culture, such that individuals with more similar genetic sequences also have more similar linguistic features. This gene–language association is shaped by processes as diverse as trade, wars, and migrations. The parallels and discontinuities found by studying the movements of people and the spread of ideas are products of both biological and cultural history (Cavalli-Sforza and Feldman 1981; Boyd and Richerson 1985; Creanza et al. 2017). Methods that explore the distributions of languages and genes can thus support or discredit purported migrations, disentangle population structure, and describe how human variation has arisen.

After the first human genetic data became available, early studies measured the relationship of genetics and language through qualitative observations of language distributions coupled with available allele frequencies (Greenberg et al. 1986; Cavalli-Sforza et al. 1988; Richards et al. 1999). Other researchers observed long-term patterns of co-evolution by observing parallels between linguistic and genetic trees (Calafell et al. 1996) or by comparing the spatial distributions of genetic and linguistic rates of change (Sokal 1988; Barbujani and Sokal 1990, 1991). This early research broadly supported Darwin's hypothesis in finding that genetic distances tended to be higher between members of different language families. These studies were limited by the availability of genetic data, often using small sequenced regions of mitochondrial DNA or statistics about blood groups. Even with limited data, researchers could identify geographically limited admixture in both genetic similarity and linguistic markers (Lansing et al. 2007). Greater availability of linguistic and genetic data allowed some researchers to compare genetic and linguistic trees directly, a process that is particularly effective when the populations have undergone well-defined demographic events such as migrations to new regions (Hunley et al. 2008; Barbieri et al. 2022). Others have leveraged increased sampling in certain geographic regions to test different models of human expansion, tentatively supporting population structuring in Native Americans (Jay et al. 2011), admixture in Central Asia (Thouzeau et al. 2017), and particular routes for Bantu migrations (González-Santos et al. 2022). The direct relationship between linguistic and genetic distances can be measured in multiple ways (Matsumae et al. 2021), but one common method is the Mantel test and partial

9780192886491_Book.indb 228 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:45

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Mantel test (Balanovsky et al. 2011; Sun et al. 2013; Thouzeau et al. 2017; Zhang et al. 2019). Mantel tests can measure relationships between sets of distance matrices and have found relationships between linguistic and genetic distances, but this type of test has disadvantages and has been discouraged in favour of other methods (Legendre et al. 2015). In particular, unless the hypothesis being tested is specifically related to distances, methods that rely on distances alone can collapse the variation present in the data and produce misleading results (Legendre and Fortin 2010; Guillot and Rousset 2013). In this chapter, we explore another method for comparing the variation in sets of high-dimensional data: Procrustes analysis.

13.2 General methods for Procrustes analyses

Procrustes analysis has been used in several recent studies to identify spatial patterns in two datasets, such as geographic structure shared by genes and languages. This method aligns two sets of data describing the same group and uses the spatial relationships between the two sets of data to measure similarity (Gower 1975; Dryden and Mardia 1998; Gower and Dijksterhuis 2004; Wang et al. 2010). It can describe general trends common to these datasets and also determine where these trends diverge. Of the two datasets used, X and Y, Procrustes analysis keeps one set of data X static, while rotating and transforming the other data Y to produce $Y_{rotated}$, minimizing the set of distances between all pairs of points x_i and $y_{i,\,rotated}$ for each sampling location or individual i (Figure 13.1). It returns $Y_{rotated}$ and uses the sum of these distances between two sets of data X and $Y_{rotated}$, known as the disparity, to quantify the similarity between the shapes of the two datasets (see Figure 13.1, *compare* step). The residuals from the rotation—the distances between x_i and $y_{i,\,rotated}$—can then be used to identify which individuals or groups are representative of, or deviate from, the general trends in the data.

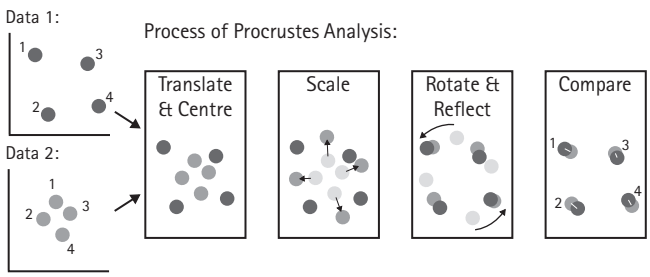


FIGURE 13.1 The steps of a Procrustes analysis to compare two sets of data.

In this example, two sets of two-dimensional data are plotted. One set of data is fixed in space, and the other set of data is translated, scaled, and rotated and/or reflected to minimize the distances between corresponding points in the comparison of the two datasets.

9780192886491_Book.indb 229 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:45

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Procrustes analysis has been used to measure the relationships between language and genetics at several scales and in many regions. When the movement of languages and of people correspond closely to one another, Procrustes correlations will be high and the disparities between the sets of data will be evenly distributed. When a group of people with a different genetic background adopts a new language, or when a group speaks the same language despite an influx of migrants and genetic admixture, Procrustes can help identify such gene–language mismatch, because the disparity between genetic and linguistic points for such mismatched populations will be greater. Researchers have applied it to measure these relationships at large, continent-wide scales, finding that genetic variation tracks both syntactic (Colonna et al. 2010) and phonemic variation (Creanza et al. 2015) across language families. At these large spatial scales, the language a population speaks is approximately as informative of their genetic variation as their location is (Creanza et al. 2015). In eastern Africa, Procrustes analysis has demonstrated that the correlation between genetics and language is decreasing as younger people in the region are shifting to more common primary languages, moving into cities, and becoming more integrated with the market (Atkinson et al. 2022). Another study of Bantu speakers in Sub-Saharan Africa used Procrustes to clarify the dynamics of the Bantu Expansion (González-Santos et al. 2022). By comparing word similarities throughout the region with individuals' genetics, they identified certain geographic regions in which Bantu-speaking populations were more likely to have both genetic and linguistic admixture from non-Bantu-speaking populations.

13.3 Preparing genetic and language data for analysis

The study of human history through gene–language analyses has become more viable because of the growing corpora of genomic data and anthropological databases. The genomic era has led to a large collection of human genetic sequences being shared online (1000 Genomes Project Consortium 2015; Mallick et al. 2016; Koenig et al. 2023) and numerous software packages for manipulating those collections of data (Danecek et al. 2011, 2021; Lawson et al. 2012; Mathe and Davis 2016). Since different parts of the genome can be inherited maternally, paternally, or from both parents, these can be used to measure the roles of sex-biased demographic processes. Similarly, databases such as the World Atlas of Language Structures (WALS; Haspelmath et al. 2005), the Atlas of Pidgin and Creole Language Structures (APiCS; Michaelis et al. 2013), PHOIBLE (Moran and McCloy 2019), Glottolog (Hammarström et al. 2022), and The Rosetta Project (2008) provide rich inventories of linguistic features across the world. These databases include morphosyntactic, lexical, and phonemic data for hundreds or thousands of languages. Together, there is a considerable number of populations for which there is an overlap of meaningful linguistic data and high-quality genetic data suited for population genetics.

9780192886491_Book.indb 230 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:45

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Procrustes analysis uses two informative, quantitative sets of data, which are often made by capturing the variation present in high-dimensional data, such as binary presence–absences of many characteristics (Figure 13.2). For genetic data, researchers often assemble sets of data where individual data points are not necessarily informative, including sets of frequencies of single-nucleotide polymorphisms, indels, or copy-number variants; after selecting which data to use, the researcher can use principal components analysis (PCA) to extract several uncorrelated vectors that describe a disproportionate amount of variance in this data (Pearson 1901). These transformed components often contain patterns that mirror the geographic distribution of individuals (Novembre et al. 2008; Jolliffe and Cadima 2016). Similarly, linguistic data can be described via PCA, particularly quantitative representations of phonetic, morphosyntactic, or word-choice data; Baayen et al. 2006, Huang et al. 2016). The use of PCA is not necessary for Procrustes analysis—any informative, quantitative data can be used—but it is an effective way of describing large sets of data in two or three axes of variation, and Procrustes analyses need these axes of variation to compare each set of data to another, particularly to geographic location data.

How a researcher selects these data plays a large role in shaping both the PCA and Procrustes analysis. To conduct quantitative analyses, both genetic and linguistic data need to be representative of variation in the region of study. For a representative sample, researchers should avoid oversampling some parts of the area; PCA is part of a family of ordination techniques that find and organize meaningful axes of variation in the data, and selecting many points that are too similar to one another emphasizes the variance within those samples, inflating their importance relative to that of other samples. Which data are collected from each site is also important. Different sorts of linguistic elements (phonemes, syntax, lexicon, etc.) have different distributions between and within languages, so care should be taken when selecting which elements to use (Plank 1998; Nichols 2003). For example, the incidence of multiple consonants, such as stops and aspirated stops, may be correlated or independent of one another in the study region; if they are correlated, then including many associated consonants as independent data points will increase the amount of variation that can be expressed as a single principal component (PC), increasing that PC's importance. For genetic data, the process of data curation has been considered in great detail. Since some elements of the genome are disproportionately inherited together because they are nearby on a chromosome—a phenomenon known as linkage disequilibrium—it is common to select genetic variants that are not closely linked to one another (Baran et al. 2013). Trimming genetic data using a linkage disequilibrium threshold produces a more accurate picture of genetic variation on the autosomes (as opposed to the Y chromosome and mtDNA, which do not undergo recombination). Both PCA and Procrustes analysis rely on linear relationships between and within data, so researchers should consider the expected distribution of variation when designing a study (Balbuena et al. 2013; Aguillon et al. 2017). While it can be useful to check whether a PCA returns meaningful results, researchers should not fine-tune the data to make the resulting PCs match a particular interpretation or expectation, as this increases the risk of a false positive.

9780192886491_Book.indb 231 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:45

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

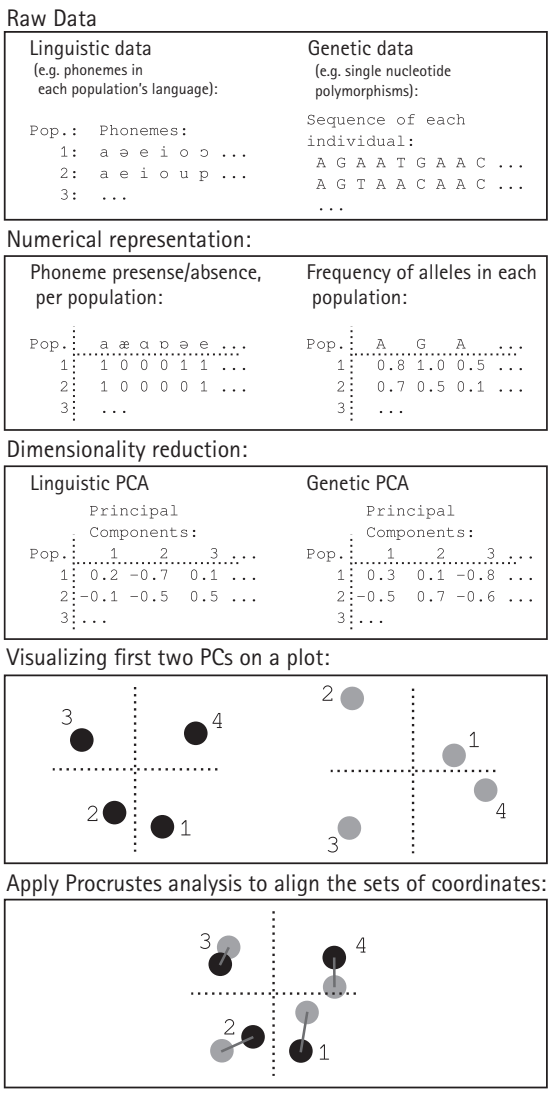


FIGURE 13.2 Schematic of the workflow from processing raw data to applying a Procrustes analysis.

For Procrustes, the raw data must be processed into quantitatively meaningful numeric matrices. Depending on the number of columns and the types of information in these data matrices, principal component analyses can reduce the dimensionality of the dataset so that two to three columns can be compared from each. These data can be visualized on a plot. In a Procrustes analysis, one set of data is then translated, scaled, and rotated to minimize the distance between corresponding points in another dataset.

Finally, researchers must select how many PCs will be used in the Procrustes analysis. To be able to compare similar data, they should select an identical number of PCs for both genetic and linguistic data. Since these are often both compared to the spatial distribution of samples, two or three PCs match the dimensionality of the

9780192886491_Book.indb 232 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:46

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

samples' spatial distribution in latitude/longitude coordinates or a projection of these coordinates onto a sphere. Note that other dimensionality reduction techniques exist, particularly multidimensional scaling methods that start from distance matrices, but we focus on PCA here and encourage the use of raw data over distance matrices, as we discuss later.

The inputs for Procrustes analysis are two matrices of the same size, for which corresponding rows represent datapoints from the same population or individual (Figure 13.2). Given two matrices representing the first n principal components for both sets of data, single functions exist to run a Procrustes analysis in MATLAB (procrustes), R (procrustes and protest in the *vegan* package), and Python (procrustes in the S*ciPy* package), among others. These functions centre and scale both matrices, and then transform the second matrix, minimizing the Procrustes sum of squared distances (*SSD*, also known as disparity) between the two matrices (Gower 1975; Dryden and Mardia 1998). The *SSD* is found by calculating the residuals between x_i and $y_{i,rotated}$ for all points i (see 'compare' step in Figures 13.1 and 13.2), squaring these distances, and finding the sum across all points. Due to scaling, the *SSD* lies between 0 (perfect match of the two datasets) and 1 (no linear transformation describes the relationship between datasets, as would occur if one or both sets of points are randomly distributed). This disparity can be converted to the Procrustes similarity statistic t_o (also known as the Procrustes correlation) via the formula: $t_o = \sqrt{(1 - SSD)}$. The Procrustes correlation has the inverse interpretation to that of the *SSD*: t_o will equal 0 when there is no relationship between the shape of the two matrices and 1 when the two have identical shapes.

Since the Procrustes algorithm minimizes the disparity between datasets, it is straightforward to parse the results. Larger disparity between sets of data, assuming identical dimensionality, means that they are less similar. To find samples that are less concordant, a researcher would find the distances between all pairs of points x_i and $y_{i,\ rotated}$; points—or groups of points—with greater distances are less representative of the general linear pattern between the datasets. These distances are the Procrustes residuals, and larger residuals denote those samples that contribute most to disparity (Colonna et al. 2010) (Figure 13.3). A common significance test for Procrustes analysis is ProTest, a method that permutes the rows of the original data thousands or millions of times and carries out Procrustes analysis to return a disparity value for each permutation (Jackson 1995). The empirical p-value is calculated as the proportion of disparities ($SSD_{permuted}$) from all permutations that have a lower disparity than that of the actual data (SSD_{actual}). For example, if none of the 10,000 Procrustes analyses of permuted data have a lower disparity than the real data, the empirical p-value is $p<10^{-4}$.

9780192886491_Book.indb 233 /12_first_proofs/first_pr.../xml_for_typesetting 27-Mar-25 23:02:46

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

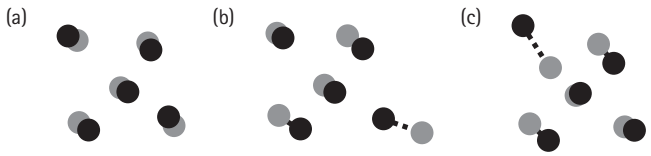


FIGURE 13.3 Illustration of residuals.

For each Procrustes-rotated set of data, we show residuals between example linguistic (black) and genetic data (grey). In the first panel (a), the linguistic and genetic data are well aligned, with similar residuals between the transformed data. In the second (b) and third (c) panels, one pair of points has a large residual, indicated by the dotted line. In one of these (b), the population's genetic distance to its neighbours is greater than its linguistic distance, suggesting that this population, which may have been isolated from the others, gained features of language similar to those of the other populations. In the final panel (c), a population may have adopted features from a distantly related language without proportional gene flow due to the movement of people.

13.4 Incorporating geography into Procrustes analyses

By comparing residuals, Procrustes analysis can be modified to jointly measure the similarity of linguistic, genetic, and geographic variation (Wang et al. 2010). Procrustes analysis can measure how geographically representative each set of data is, and Procrustes rotations have been leveraged to visualize these relationships on a map (Creanza et al. 2015) (Figure 13.4). This can be done via directly transforming each of the genetic and linguistic datapoints onto latitude and longitude, but at larger distances and farther from the equator, this is likely to distort points. This type of map projection,

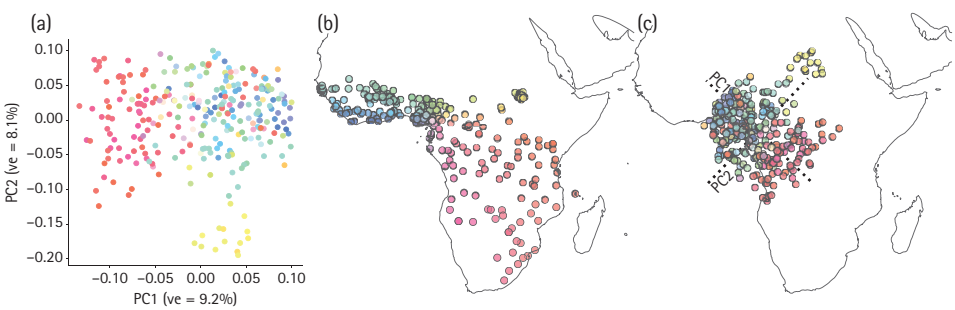


FIGURE 13.4 A Procrustes rotation of the first two principal components of phonemic variation among Niger-Congo languages.

We used a presence/absence matrix of phoneme inventories analysed in Creanza et al. (2015) as a measure of linguistic variation and then conducted PCA on the 279 Niger-Congo languages in that database. We plotted the first two principal components of these points (A) and coloured them according to their location, mapped in panel (B). We then used Procrustes analysis to rotate the linguistic PCs to best match the languages' locations and plotted these Procrustes-rotated PCs on a map (C).

9780192886491_Book.indb 234 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:47

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

in which latitude and longitude are plotted in a simple grid with constant spacing, is called equirectangular projection; by including an outline of an equirectangular map with the geographically projected data, researchers can at least give a visual indication of the resulting distortions. To better account for the shape of the Earth, latitudes and longitudes could instead be converted into points in three-dimensional Cartesian space, in which distances between points are Euclidean rather than great circle (Atkinson et al. 2022); the differences between the two distance metrics are meaningful at only the largest spatial scales, and the transformed matrix can be projected onto the geographic coordinate system. Alternatively, one could find the great-circle distance between sampling locations and use principal coordinate analysis to find an appropriate two-dimensional representation of the data, though this makes interpretation of the transformed matrix more difficult.

In addition to comparing both language and genetics to spatial distributions, researchers can use a Partial ProTest, which takes advantage of the residuals from one Procrustes analysis to measure another relationship (Peres-Neto and Jackson 2001). For example, the residuals from a Procrustes analysis of geography on genetics represent the variation in genetics that is not linearly associated with spatial distances; these residual vectors—which have the same dimensionality as the original data—measure the association of genetic variation and phoneme inventories after controlling for geography. In these ways, a single Procrustes analysis of two matrices can directly compare genetics, language, and geography, probing the relationship between them and allowing for more controlled analyses of their interactions.

13.5 Advantages and drawbacks

Procrustes analysis is a versatile tool that offers some advantages over other methods of measuring covariation. The flexibility of Procrustes allows researchers to use raw or dimensionally reduced data instead of—or in addition to—distance matrices. As more data become accessible and researchers are not restricted to using distance matrices, the use of purely distance-based approaches such as Mantel tests has been disfavoured, particularly if the study question can be phrased in such a way that it does not refer to distances (Legendre and Fortin 2010; Guillot and Rousset 2013). Methods that preserve the variation in the original data, such as Procrustes or regression analyses, are more likely to identify genuine correlations among variables and are easier to interpret (Peres-Neto and Jackson 2001; Legendre et al. 2015). Unless researchers are explicitly interested in the relationships of sets of distances or variance–covariance matrices, creating these from the original data makes it more difficult to interpret both data and results (Novembre and Stephens 2008). Since Procrustes analysis is both more sensitive than the Mantel test and can control for other data (similarly to partial or hierarchical

9780192886491_Book.indb 235 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:47

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Mantel tests; Peres-Neto and Jackson 2001) while avoiding the issues of a distance-based approach, it is a powerful method for gene–language analyses.

There are several drawbacks to Procrustes analyses, and these should be taken into account before choosing it for a study. Both language and genetics have been shaped by evolutionary forces that can create specific patterns of covariance in the data (Hansen and Martins 1996), and because Procrustes is not a specifically evolutionary method, it does not explicitly make use of, or test for, this covariation. Since genetic and linguistic variation between human populations often cannot be represented though phylogenetic trees or similar networks, for example, in the case of admixture, this may not be a concern. For studies on large scales where phylogenetic relationships are most relevant, there are extensions of Procrustes analysis that rely on (Openshaw and Keogh 2014) and test (Balbuena et al. 2013) explicit models of evolution using given networks of relationships between populations. In addition, reducing the dimensionality of one or both sets of data (genetic or linguistic) comes with costs to interpretability and repeatability (Uyeda et al. 2015). Despite these drawbacks, Procrustes analysis can describe broad patterns between datasets and identify outlying points, making it an effective analytical tool. In particular, understanding the distribution of these outliers and the relationship between datasets can suggest which regions should be sampled for future studies. It is a method that can be adapted to exploring the interactions between many forms of data, and its results are a starting point to a more thorough and precise study of gene–language covariation and the processes that lead to it.

13.6 Data availability

See https://github.com/CreanzaLab/procrustes_genelang_example for code to do a basic Procrustes analysis to generate the plots in Figure 13.4.

References

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., et al. (2015). 'A global reference for human genetic variation', *Nature* 526: 68–74.

Aguillon, S. M., Fitzpatrick, J. W., Bowman, R., et al. (2017). 'Deconstructing isolation-by-distance: The genomic consequences of limited dispersal', *PLoS Genetics* 13: e1006911.

Atkinson, E. G., Dalvie, S., Pichkar, Y., et al. (2022). 'Genetic structure correlates with ethnolinguistic diversity in eastern and southern Africa', *American Journal of Human Genetics* 109: 1667–1679.

Baayen, R. H., Feldman, L. B., and Schreuder, R. (2006). 'Morphological influences on the recognition of monosyllabic monomorphemic words', *Journal of Memory and Language* 55: 290–313.

Balanovsky, O., Dibirova, K., Dybo, A., et al. (2011). 'Parallel evolution of genes and languages in the Caucasus region', *Molecular Biological Evolution* 28: 2905–2920.

9780192886491_Book.indb 236 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:47

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Balbuena, J. A., Míguez-Lozano, R., and Blasco-Costa, I. (2013). 'PACo: A novel procrustes application to cophylogenetic analysis', *PLoS One* 8: e61048.

Baran, Y., Quintela, I., Carracedo, A., et al. (2013). 'Enhanced localization of genetic samples through linkage-disequilibrium correction', *American Journal of Human Genetics* 92: 882–894.

Barbieri, C., Blasi, D. E., Arango-Isaza, E., et al. (2022). 'A global analysis of matches and mismatches between human genetic and linguistic histories', *Proceedings of the National Academy of Sciences of the United States of America* 119: e2122084119.

Barbujani, G. and Sokal, R. R. (1990). 'Zones of sharp genetic change in Europe are also linguistic boundaries', *Proceedings of the National Academy of Sciences of the United States of America* 87: 1816–1819.

Barbujani, G. and Sokal, R. R. (1991). 'Genetic population structure of Italy: II. Physical and cultural barriers to gene flow', *American Journal of Human Genetics* 48: 398–411.

Boyd, R. and Richerson, P. J. (1985). *Culture and the Evolutionary Process*. Chicago, IL: University of Chicago Press.

Calafell, F., Bertranpetit, J., Rendine, S., et al. (1996). 'Population history of Corsica: A linguistic and genetic analysis', *Annual of Human Biology* 23: 237–251.

Cavalli-Sforza, L. L. and Feldman, M. W. (1981). 'Cultural transmission and evolution: A quantitative approach', *Monogram of Popular Biology* 16: 1–388.

Cavalli-Sforza, L. L., Piazza, A., Menozzi, P., et al. (1988). 'Reconstruction of human evolution: Bringing together genetic, archaeological, and linguistic data', *Proceedings of the National Academy of Sciences of the United States of America* 85: 6002–6006.

Cavalli-Sforza, L. L., Minch, E., and Mountain, J. L. (1992). 'Coevolution of genes and languages revisited', *Proceedings of the National Academy of Sciences of the United States of America* 89: 5620–5624.

Colonna, V., Boattini, A., Guardiano, C., et al. (2010). 'Long-range comparison between genes and languages based on syntactic distances', *Human Heredity* 70: 245–254.

Creanza, N., Ruhlen, M., Pemberton, T. J., et al. (2015). 'A comparison of worldwide phonemic and genetic variation in human populations', *Proceedings of the National Academy of Sciences of the United States of America* 112: 1265–1272.

Creanza, N., Kolodny, O., and Feldman, M. W. (2017). 'Cultural evolutionary theory: How culture evolves and why it matters', *Proceedings of the National Academy of Sciences of the United States of America* 114: 7782–7789.

Danecek, P., Auton, A., Abecasis, G., et al. (2011). 'The variant call format and VCFtools', *Bioinformatics* 27: 2156–2158.

Danecek, P., Bonfield, J. K., Liddle, J., et al. (2021). 'Twelve years of SAMtools and BCFtools', *Gigascience* 10, DOI: 10.1093/gigascience/giab008.

Darwin, C. (1871). *The Descent of Man*. London: John Murray.

Di Corcia, T., Sanchez Mellado, C., Davila Francia, T. J., et al. (2017). 'East of the Andes: The genetic profile of the Peruvian Amazon populations', *American Journal of Physical Anthropology* 163: 328–338.

Dryden, I. L. and Mardia, K. V. (1998). *Statistical Shape Analysis*. Chichester: John Wiley & Sons.

González-Santos, M., Montinaro, F., Grollemund, R., et al. (2022). 'Exploring the relationships between genetic, linguistic and geographic distances in Bantu-speaking populations', *American Journal of Biological Anthropology* 179: 104–117.

Gower, J. C. (1975). 'Generalized procrustes analysis', *Psychometrika* 40: 33–51.

9780192886491_Book.indb 237 /12_first_proofs/first_pr…/xml_for_typesetting 27-Mar-25 23:02:47

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Gower, J. C. and Dijksterhuis, G. B. (2004). *Procrustes Problems*, DOI: 10.1093/acprof:oso/9780198510581.001.0001.

Greenberg, J. H., Turner, C. G., Zegura, S. L., et al. (1986). 'The settlement of the Americas: A comparison of the linguistic, dental, and genetic evidence [and comments and reply]', *Current Anthropology* 27: 477–497.

Guglielmino, C. R., Viganotti, C., Hewlett, B., et al. (1995). 'Cultural variation in Africa: Role of mechanisms of transmission and adaptation', *Proceedings of the National Academy of Sciences of the United States of America* 92: 7585–7589.

Guillot, G. and Rousset, F. (2013). 'Dismantling the mantel tests', *Methods in Ecological Evolution* 4: 336–344.

Haber, M., Gauguier, D., Youhanna, S., et al. (2013). 'Genome-wide diversity in the levant reveals recent structuring by culture', *PLoS Genetics* 9: e1003316.

Hammarström, H., Forkel, R., Haspelmath, M., et al. (2022). glottolog/glottolog: Glottolog database 4.7., DOI: 10.5281/ZENODO.7398962.

Hansen, T. F. and Martins, E. P. (1996). 'Translating between microevolutionary process and macroevolutionary patterns: The correlation structure of interspecific data', *Evolution* 50: 1404–1417.

Haspelmath, M., Dryer, M. S., Gil, D., et al. (2005). *The World Atlas of Language Structures*. Oxford: Oxford University Press.

Huang, Y., Guo, D., Kasakoff, A., et al. (2016). 'Understanding U.S. regional linguistic variation with Twitter data analysis', *Computers, Environment and Urban Systems* 59: 244–255.

Hunley, K., Dunn, M., Lindström, E., et al. (2008). 'Genetic and linguistic coevolution in Northern Island Melanesia', *PLoS Genetics* 4: e1000239.

Jackson, D. A. (1995). 'PROTEST: A PROcrustean Randomization TEST of community environment concordance', *Écoscience* 2: 297–303.

Jay, F., François, O., and Blum, M. G. B. (2011). 'Predictions of native American population structure using linguistic covariates in a hidden regression framework', *PLoS One* 6: e16227.

Jolliffe, I. T. and Cadima, J. (2016). 'Principal component analysis: A review and recent developments', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374: 20150202.

Koenig, Z., Yohannes, M. T., Nkambule, L. L., et al. (2023). 'A harmonized public resource of deeply sequenced diverse human genomes', *bioRxiv*, DOI: 10.1101/2023.01.23.525248.

Lansing, J. S., Cox, M. P., Downey, S. S., et al. (2007). 'Coevolution of languages and genes on the island of Sumba, eastern Indonesia', *Proceedings of the National Academy of Sciences of the United States of America* 104: 16022–16026.

Lawson, D. J., Hellenthal, G., Myers, S., et al. (2012). 'Inference of population structure using dense haplotype data', *PLoS Genetics* 8: e1002453.

Legendre, P. and Fortin, M.-J. (2010). 'Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data', *Molecular Ecology Resources* 10: 831–844.

Legendre, P., Fortin, M.-J., and Borcard, D. (2015). 'Should the Mantel test be used in spatial analysis?', *Methods in Ecological Evolution* 6: 1239–1247.

Lumsden, C. J. and Wilson, E. O. (1981). 'Genes, mind, and ideology', *The Sciences* 21: 6–8.

Mallick, S., Li, H., Lipson, M., et al. (2016). 'The Simons Genome Diversity Project: 300 genomes from 142 diverse populations', *Nature* 538: 201–206.

Mathe, E. and Davis, S. D. (eds.) (2016). *Statistical Genomics: Methods and Protocols*. 1st ed. New York: Humana Press.

9780192886491_Book.indb 238 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:47

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Matsumae, H., Ranacher, P., Savage, P. E., et al. (2021). 'Exploring correlations in genetic and cultural variation across language families in northeast Asia', *Scientific Advances* 7, DOI: 10.1126/sciadv.abd9223.

Michaelis, S. M., Maurer, P., Haspelmath, M., et al. (2013). The Atlas of Pidgin and creole language structures online.

Moran, S. and McCloy, D. (2019). PHOIBLE 2.0. *Jena: Max Planck Institute for the Science of Human History*.

Nichols, J. (2003). 'Diversity and stability in language', in B. D. Joseph and R. D. Janda (eds.), *The Handbook of Historical Linguistics*, 283–310. Hoboken, NJ: Blackwell Publishing.

Novembre, J. and Stephens, M. (2008). 'Interpreting principal component analyses of spatial population genetic variation', *Nature Genetics* 40: 646–649.

Novembre, J., Johnson, T., Bryc, K., et al. (2008). 'Genes mirror geography within Europe', *Nature* 456: 98–101.

Openshaw, G. H. and Keogh, J. S. (2014). 'Head shape evolution in monitor lizards (Varanus): Interactions between extreme size disparity, phylogeny and ecology', *Journal of Evolutionary Biology* 27: 363–373.

Pearson, K. (1901). 'Principal components analysis', *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 6: 559.

Peres-Neto, P. R. and Jackson, D. A. (2001). 'How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test', *Oecologia* 129: 169–178.

Plank, F. (1998). 'The co-variation of phonology with morphology and syntax: A hopeful history', *Linguistic Typology* 2, DOI: 10.1515/lity.1998.2.2.195.

Rickards, O., Martínez-Labarga, C., Lum, J. K., et al. (1999). 'mtDNA history of the Cayapa Amerinds of Ecuador: Detection of additional founding lineages for the Native American populations', *American Journal of Human Genetics* 65: 519–530.

The Rosetta Project. (2008). rosettaproject.org.

Sokal, R. R. (1988). 'Genetic, geographic, and linguistic distances in Europe', *Proceedings of the National Academy of Sciences of the United States of America* 85: 1722–1726.

Sun, H., Zhou, C., Huang, X., et al. (2013). 'Correlation between the linguistic affinity and genetic diversity of Chinese ethnic groups', *Journal of Human Genetics* 58: 686–693.

Thouzeau, V., Mennecier, P., Verdu, P., et al. (2017). 'Genetic and linguistic histories in Central Asia inferred using approximate Bayesian computations', *Proceedings of the Royal Society B: Biological Sciences* 284, DOI: 10.1098/rspb.2017.0706.

Uyeda, J. C., Caetano, D. S., and Pennell, M. W. (2015). 'Comparative analysis of principal components can be misleading', *Systematic Biology* 64: 677–689.

Wang, C., Szpiech, Z. A., Degnan, J. H., et al. (2010). 'Comparing spatial maps of human population-genetic variation using Procrustes analysis', *Statistical Applications in Genetics and Molecular Biology* 9: Article 13.

Zhang, M., Zheng, H.-X., Yan, S., et al. (2019). 'Reconciling the father tongue and mother tongue hypotheses in Indo-European populations', *National Science Review* 6: 293–300.

9780192886491_Book.indb 239 /12_first_proofs/first_pr[illegible]/xml_for_typesetting 27-Mar-25 23:02:47

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

9780192886491_Book.indb 240 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:47

CHAPTER 14

STUDYING LONG-TERM EVOLUTIONARY PROCESSES OVER SHORT TIMESCALES

Methods in cumulative cultural evolution

MAXIME DEREX AND CHARLOTTE BRAND

14.1 Introduction

Complex cultural traits have allowed humans to settle in habitats for which they are poorly suited genetically (Boyd and Richerson 1985; Boyd et al. 2011; Henrich 2015). Bows, kayaks, spears, and harpoons are only a few examples among the myriad technologies that sustain humans in almost every terrestrial environment on earth. These finely tuned technologies are not produced in isolation by especially gifted individuals, but result from a cumulative cultural evolutionary process in which innovations are gradually accumulated across many generations (Boyd and Richerson 1996; Boyd et al. 2011; Derex et al. 2019).

Social learning is critical to cumulative cultural evolution because it allows innovations to be transmitted between individuals and across generations. Yet, the ability to learn socially appears to be widespread in animals, while the accumulation of cultural innovations is not (Boyd and Richerson 1996). Moreover, while there is a general trend towards richer and more complex cultural repertoires in humans, it appears that cultural complexity does not increase steadily and monotonically over time. Periods of both sudden cultural accumulation and cultural regression have been documented (Henrich 2004; d'Errico and Stringer 2011; Riede 2014). These observations suggest that cultural accumulation is not a trivial process but occurs only when very specific conditions are met.

Theoretical models have long been used to study the conditions that are conducive to cumulative cultural evolution (Cavalli-Sforza and Feldman 1981; Boyd and Richerson

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

1985). More recently, cultural evolution researchers have used experiments to investigate how beneficial modifications are selectively preserved and accumulated over successive generations (Caldwell and Millen 2008b; Caldwell et al. 2016). Cumulative culture requires the production of innovations and their propagation within social groups. Thus, experiments that are rigorously designed to track the learning, transmission, and modifications of innovations can shed light on the underlying mechanisms that affect cultural accumulation. These experiments allow drastic compression of the 'evolutionary' timescale and have proved successful in addressing a wide range of questions concerning the production, transmission, and maintenance of cultural traits (Caldwell and Millen 2008a; Beppu and Griffiths 2009; Derex, Beugin, et al. 2013; Caldwell et al. 2019; Derex et al. 2019; Lucas et al. 2020; Thompson et al. 2022). While early experiments mostly focused on identifying the minimal conditions allowing cultural information to accumulate, recent experiments have started to explore broader factors that affect the dynamics of cumulative cultural evolution. For instance, experimental results indicate that, perhaps counterintuitively, reducing group connectedness can result in higher levels of cultural accumulation (Derex and Boyd 2016).

Surprisingly, the relationship between different forms of communication and cumulative cultural evolution has remained relatively understudied in the cultural evolutionary literature. Indeed, despite recent calls for better integration of communication into cultural evolution experiments (Brand et al. 2021; Singh et al. 2021), experimental studies have only begun to scratch the surface of the complex relationship between communication and cultural evolution. So far, experiments have mostly explored how some forms of pedagogy, such as gestural and verbal teaching, affects the stability of cultural information (e.g. Morgan et al. 2015; Lucas et al. 2020).

In this chapter, we present the experimental methodology typically used to study cumulative cultural evolution in the lab (for an introduction/review of individual-based models of Cultural Evolution, see Acerbi et al. 2022). We highlight the limitations and challenges associated with this method, discuss how human language has thus far been addressed within cultural evolution experiments, and suggest potential fruitful avenues to test so far untested hypotheses about the relationship between different forms of communication and cultural evolution.

14.2 Key principles and assumptions

A few basic requirements must be met in order to experimentally study cumulative cultural evolution in the lab. For the experimenter, the goal is to create a setup in which the necessary and sufficient conditions for the gradual improvement of cultural traits are met. These can be summarized as four core criteria: (i) a change in behaviour; (ii) transmission of this change via social learning; (iii) an improvement in performance; and (iv) the sequential repetition of the first three criteria (Mesoudi and Thornton 2018). It is important to note that, depending on the specific question being addressed,

9780192886491_Book.indb 242 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:47

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

additional criteria may be considered. Indeed, while the conjunction of the aforementioned criteria leads to gradual improvement across generations of learners, they may not comprehensively account for all the facets of human cumulative culture (Derex 2021). For instance, some have argued that human cumulative culture is characterized by the presence of increasingly complex and harder-to-learn cultural traits (which seems to distinguish it from animals' cultural repertoires that may be composed of multiple but not increasingly complex traits; Dean et al. 2013). Thus, criteria such as functional dependence (where an improvement is functionally dependent on a previous one) or recombination (where a new trait results from the combination of existing traits) are sometimes considered in cumulative cultural evolution experiments (for recent discussions about what constitutes cumulative cultural evolution, see Mesoudi and Thornton 2018; Miton and Charbonneau 2018; Derex 2021).

Regardless of the criteria considered, special attention must be given to both the task and the conditions under which participants will interact to ensure that the conditions are conducive to cumulative cultural evolution. We go through these principles in detail below.

14.2.1 Choosing the task

What makes a task appropriate to study cumulative cultural evolution ultimately depends on the question at hand. This means that your research question should determine the task on which your experiment will be based, not the other way around. Nevertheless, a few core principles are useful at the design stage.

The nature of the task and its goal are not inherently important and can take various forms. Cultural evolution experiments have relied on tasks as diverse as paper aeroplane building, knot-tying, stone tool making, totem pole building, among many others (Caldwell and Millen 2008a; Muthukrishna et al. 2014; Derex and Boyd 2015; Morgan et al. 2015). Experiments have also relied on physical as well as computer-based tasks. What is critical is that the task in question can be solved with varying degrees of success, and, ideally, that variation in success can be evaluated easily and objectively by the experimenter. The task should be difficult enough that participants cannot solve it in a few trials, but easy enough that collective improvements can realistically be expected during the relatively short duration of an experiment. This usually requires the experimenter to pilot the experiment to ensure that the task is in the right difficulty range. For instance, all else being equal, experiments involving sizable groups require more difficult tasks than experiments involving smaller groups to prevent group performance from plateauing before the end of the experiment.

The difficulty with which participants solve the task also depends on how familiar the task is to participants. Ideally, the task should be unfamiliar to prevent participants from relying on previously acquired knowledge to solve it. Using tasks that are too familiar can make it more challenging to observe differences in performance over time because participants with previous experience with the task will perform better than

9780192886491_Book.indb 243 /12_first_proofs/first_pr…/xml_for_typesetting 27-Mar-25 23:02:47

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

truly naive individuals. This will shift the baseline performance up and will reduce the amount of variation that can be observed during the course of the experiment. Moreover, previously experienced participants will add unwanted noise to the experiment. In a transmission-chain design, for instance, experienced participants who already have their own established ways of solving the task may appear at the beginning, end, or middle of an experimental chain and disrupt the process of cultural accumulation by disregarding social information.

14.2.2 Physical vs. computer-based tasks

As mentioned above, cultural evolution experiments can be both physical and computer-based, both being common in the literature (e.g. Derex and Boyd 2015; Morgan et al. 2015). Computer-based tasks are often more convenient because they permit rigorous manipulations of payoff structures and can be administered quickly and easily to large numbers of participants. Physical tasks tend to make data collection more time-consuming, yet they feature realistic physical principles that were arguably more relevant during our evolutionary history (more 'ecologically valid'). For instance, physical tasks rely on individuals' sensorimotor skills (such as accurately striking a core in a stone tool-making experiment), or understanding of 'folk-physics', which some would argue are more relevant when trying to understand the cognitive biases and reasoning abilities of our evolutionary ancestors. In contrast, whilst computerized tasks can tap into visuo-audio perception abilities, they are often solved by relying on basic motor actions (such as pointing and clicking) which participants already master before taking part in the experiment. This means that physical tasks are more amenable to study of the transmission of skills (as opposed to knowledge, or perceptual biases) where face-to-face interaction and gestural demonstration may be critical.

Still another difference between the two types of tasks is the experimental environment in which they can be deployed. Computerized tasks can easily be deployed in the lab and online. Online recruitment platforms, such as Prolific, allow the recruitment of genuinely diverse and representative samples compared to what is often readily available to researchers, allowing samples to be broader than the usual 'university undergrad' default. The possibility to reach a more diverse and geographically distributed sample of participants, combined with the possibility to collect a large sample in a cost-effective manner, can allow for greater generalizability of results. Online tasks, however, cannot be used with groups who have limited access to digital technologies in the first place.

Ultimately, the choice to use computerized or physical tasks to study cultural evolution will depend on the specifics of the experimenters' questions and research objectives. If the research question pertains to very abstract, generalizable aspects of cultural transmission, then there is no reason why a computerized task (that is also fun and engaging for participants!) cannot capture the core criteria mentioned above and reveal insights into how people learn and transmit cultural information. If the research question relates

9780192886491_Book.indb 244 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:47

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

to whether prestigious community members make better teachers than non-prestigious members in the context of norms or skills specific to a given community, then of course using a physical task relevant for this community is likely to be more revealing.

14.2.3 Implementing transmission

For cultural evolution to happen, the experimental setup must be conducive to the transmission of information between participants. This information, however, can take different forms and be more or less useful to the learner. Experiments related to the debate about the role of high-fidelity social learning in cumulative cultural evolution, for instance, typically implement a number of experimental treatments whose difference lies in the type of cultural transmission involved (e.g. Morgan et al. 2015). Reverse engineering, for instance, involves replicating an outcome without being exposed to the details of how the outcome was achieved, while imitation provides the learner with the specific details of how the outcome was achieved. Still another treatment could involve teaching (either verbal or gestural) in which case the learner would be actively taught the details of how the outcome was achieved.

Although cultural evolution experiments typically focus on one or several of those types of transmission, options for the experimenter are virtually endless. For instance, one might think of comparing a treatment where participants transmit written information using alphanumeric characters to another treatment where participants transmit information using emojis. In experiments investigating the role of variables other than transmission mechanisms, only one transmission mechanism is usually implemented (e.g. Derex, Beugin, et al. 2013). An experimenter investigating the effect of group size on cultural accumulation, for instance, might want to compare the performance of individuals who are part of groups of two, four, and six and decide that participants from all treatments will learn by being provided with each other's outcome without being exposed to the details of how the outcome was achieved.

14.2.4 Transmission-chain vs. closed-group methods

Another choice that the experimenter must make concerns the experimental setting within which cultural transmission will take place. Different experimental settings have been used to study cumulative cultural evolution in the lab.

One of the most widely used methods is the transmission chain, in which information (e.g. skills, text, images, stories, songs) is transmitted from one generation to the next (e.g. Caldwell and Millen 2008a; Derex and Boyd 2015; Morgan et al. 2015; Derex et al. 2019). By analysing the changes that occur within the material as it is transmitted from person to person, researchers can infer the operation of systematic biases in cultural transmission, such as the effects of memory, attention, or communication (see Tamariz and Papa on iterated learning, Chapter 1, this volume). In cumulative cultural evolution

9780192886491_Book.indb 245 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:47

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

experiments, often first-generation participants are asked to solve the task without any input. Their solution is then passed on to the next participant in the chain. Within that setting, naive, first-generation participants provide the baseline performance against which the performance of subsequent participants can be compared. The researcher can then study how solutions evolve and test whether they become increasingly efficient over time. One limitation of this method is that transmission chains are prone to cultural loss because the process of cultural accumulation depends on a single individual at each generation. This means that discontinuity can be caused by individuals who, for some reason, ignore, forget, or misinterpret social information.

An alternative methodology is the closed-group method in which a group of individuals is brought together and repeatedly engages in a task over the course of the experiment (e.g. Mesoudi 2011; Derex and Boyd 2015, 2016). This method is often used in experiments where researchers want to study social learning strategies, or how some variables affect groups' or individuals' success. Compared to transmission chains, closed groups offer participants the opportunity to learn from multiple cultural models. This offers participants the opportunity to select their cultural demonstrator based on cues such as score or prestige (e.g. Mesoudi 2011; Atkisson et al. 2012; Chudek et al. 2012; Brand et al. 2020). In some experiments, participants can simultaneously learn and combine information from multiple demonstrators (Derex, Beugin, et al. 2013; Kempe and Mesoudi 2014; Muthukrishna et al. 2014). The closed-group method is well suited to computerized tasks. Indeed, with computerized tasks, it is possible to store all the decisions taken by a player on a server, which allows participants to access their other group members' solutions in real time. An individual learning condition, in which participants engage in the same task but with no social interaction, provides a baseline with which to compare group performance (Mesoudi and Whiten 2008). These experiments tend to be less time-consuming as multiple participants engage with the task at the same time, and are not dependent on the problem of simultaneous recruitment that transmission-chain designs require.

14.3 Examples

In this section, we will walk through the methodology, including data collection and analyses, that were used in two cultural evolution experiments.

14.3.1 Example 1: In-person transmission chains (Morgan et al. 2015)

14.3.1.1 Question and task

Morgan et al. used an in-person transmission-chain design to test the relative success of five social learning mechanisms to transmit stone-knapping techniques across multiple

9780192886491_Book.indb 246 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:47

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

transmission events. The authors used a task that has been of critical importance in our evolutionary past: stone tool-making.

14.3.1.2 Treatments and procedure

The task required adult human participants to learn to produce stone flakes by striking a cobble core with a hammerstone (Figure 14.1a). Each participant learnt from the previous participant in the transmission chain. Experimenters trained in stone knapping acted as demonstrator to the first participant in the chain.

Participants were randomly allocated to one of five treatments (Figure 14.1b–f): Reverse Engineering (b)—learners were provided with the flakes produced by their demonstrator but could not see the manufacturing process; Imitation (c)—learners could observe their demonstrator making flakes but could not interact with them; Basic Teaching (d)—demonstrators could manually shape the learner's grasp of their material, slow their own actions, and reorient themselves to allow the learner a clear view; Gestural Teaching (e)—learners and demonstrators could interact using gestures but could not talk to each other; Verbal Teaching (f)—learners and demonstrators were permitted to speak.

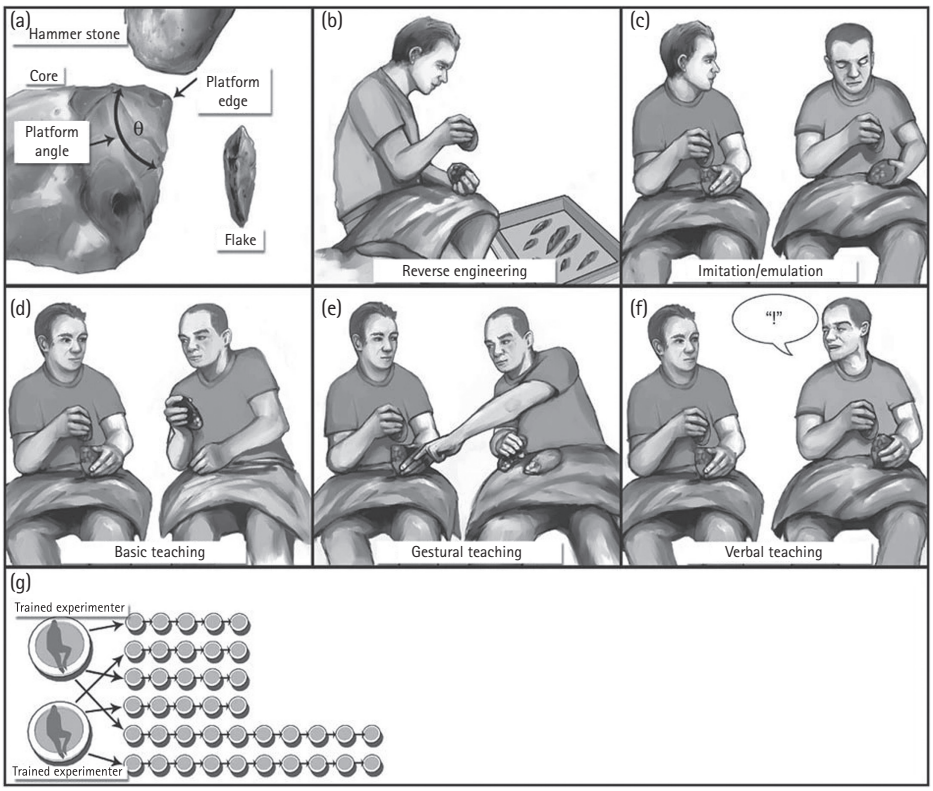


FIGURE 14.1 Diagram of Morgan et al's flint-knapping transmission-chain design.

Source: taken from https://doi.org/10.1038/ncomms7029 with permission (CC BY 4.0).

9780192886491_Book.indb 247 /12_first_proofs/first_pr…/xml_for_typesetting 27-Mar-25 23:02:48

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

The learning/teaching period lasted for five minutes. The measure of success was good-quality flakes made from a single core in a twenty-minute practice period. To ensure participant motivation, participants were paid according to their performance. To make sure demonstrators were motivated to teach effectively in the teaching conditions, participants' payments also depended on the performance of their pupil.

14.3.1.3 Data and analyses

Morgan et al. analysed 6,214 pieces of flint greater than 2 cm in diameter. All of these pieces were weighed, measured, and assessed for viability and quality by human coders. The reliability of flake viability ratings was ensured by double and triple coding by independent raters. Six different measures of individual performance were modelled: (i) the number of viable flakes produced; (ii) the total quality of flakes produced; (iii) the proportion of flakes that were viable; (iv) the rate at which viable flakes were produced; (v) the probability of a viable flake per hit; and (vi) the proportion of their core successfully reduced. These measures were modelled as a function of condition, position along the chain, interactions between condition and position, initial core mass, and random repeat-level effects.

14.3.1.4 Main result

Results revealed that participants who were taught, as opposed to learning via passive observation, produced more tools, did so more quickly, and made more efficient use of raw materials. These benefits were further enhanced by verbal, as opposed to gestural, teaching.

14.3.2 Example 2: Computer-based closed groups (Derex and Boyd 2015)

14.3.2.1 Question and task

Derex and Boyd used a computer-based experiment to investigate how social learning mechanisms and population size and structure affect the production of a complex virtual artefact (in this case, virtual totem poles; Figure 14.2). In this task, players were provided six initial basic resources that had to be combined to produce increasingly complex innovations, and the production of these complex innovations depended on the discovery of lower-level innovations (Figure 14.2). Thus, in comparison to the stone tool-making task described above, the totem task entails features such as *recombination* (where new traits result from the combination of existing traits) and *functional dependence* (where an improvement is functionally or sequentially dependent on a previous one).

14.3.2.2 Treatments and procedure

Participants sat at physically separated, networked computers and were randomly assigned to one of five treatments: individual learning treatment (1)—participants were learning in isolation and provided the baseline performance against which the

9780192886491_Book.indb 248 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:48

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

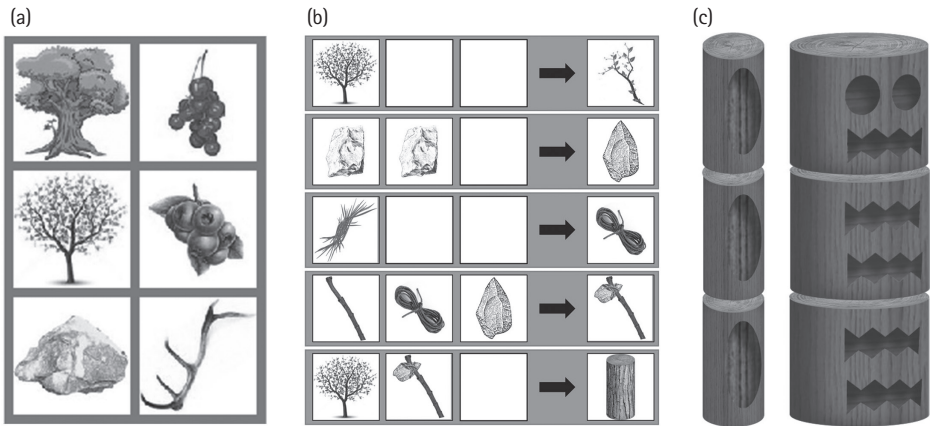


FIGURE 14.2 Experimental task of Derex and Boyd (2015) totem pole experiment.

The game simulates the real-world innovation process in which the production of complex artefacts (that is, virtual 'totem poles') depends on the discovery of high-level innovations (such as axes), which in turn is contingent on the discovery of lower-level innovations (such as stone tools), both low- and high-level innovations resulting from a specific production process. (a) The 'resources panel'—players were provided six initial basic resources that could be combined using a workshop panel containing four slots (Figure 14.3). (b) Examples of successful combinations—by placing items into a workshop panel (squares on the left; only three are depicted here), participants could produce innovations (squares on the right). Low-level innovations (created by combining basic resources) could be combined to produce higher-level innovations. Further accumulation of innovations could produce complex tools (such as axes) that potentially allowed players to get logs (by cutting trees) to build their totem. (c) Examples of totem poles—other high-level innovations (such as carving tools or pigments) could be subsequently used to refine totems to increase their value. Players' gain depended on the number of innovations they discovered and the value of their totem.

Source: adapted from https://doi.org/10.1038/ncomms9398 with permission (CC BY 4.0).

performance of social learners was compared; full social information/large group (2)—participants were part of groups of six and could learn both the innovations and the associated production processes discovered by their five other group members (example in Figure 14.3); partial social information/large group (3)—participants were part of groups of six and could learn the innovations discovered by their five other group members but could not observe the associated production processes; full social information/small groups (4)—participants were part of groups of three and could learn both the innovations and the associated production processes discovered by their two other group members; full social information/partially connected groups (5)—participants were part of a metapopulation of three groups of two participants whose connectedness patterns changed over time. In this final treatment, participants could learn both the innovations and the associated production processes discovered by whoever was part of their subgroup at the time. The experiment lasted forty-five minutes, after which subjects received a reward according to their performance.

14.3.2.3 Data and analyses

Derex and Boyd analysed participants' total score which was made up of the score of their totem (if any) plus a fixed number of points per innovation discovered. To test the effect of variables such as group size and group connectedness, analyses were run

9780192886491_Book.indb 249 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:48

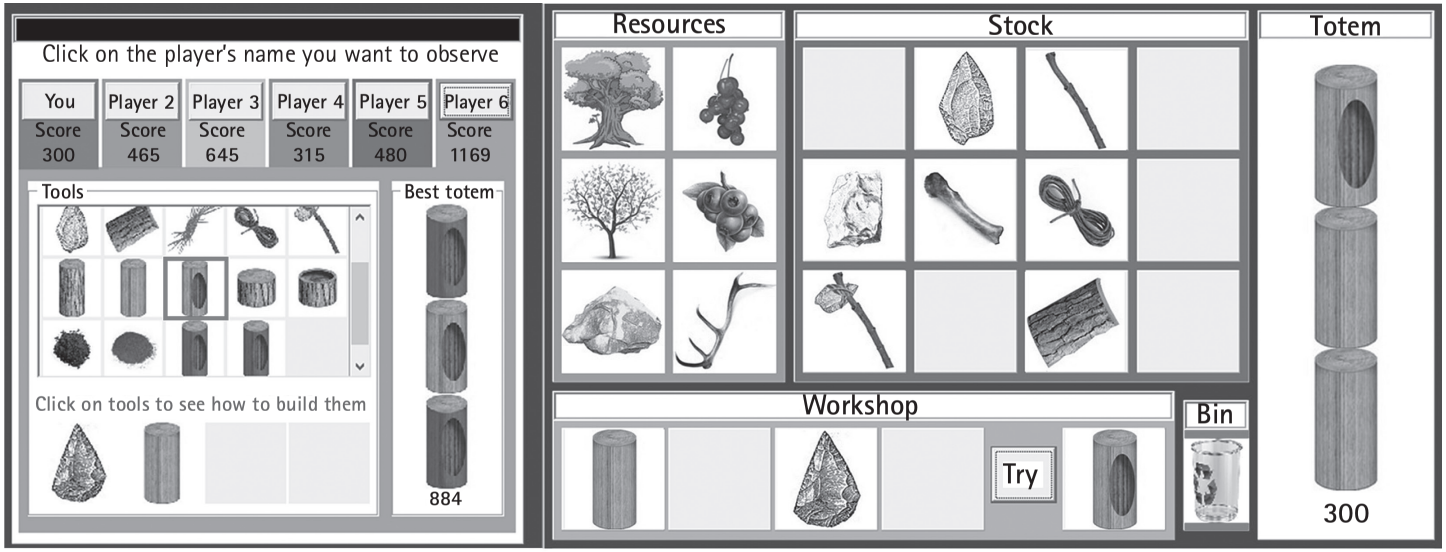


FIGURE 14.3 Game interface.

Resources could be dropped into the 'workshop panel' to be refined. Players could trigger an automatic refining process by clicking on the 'try' button. Successful combinations resulted in a new item that could be dropped into the 'stock panel' or in the 'workshop panel' to be further refined. Logs were the minimal elements that could be dropped into the 'totem panel' and provided players with a totem score. The panel on the left provided players with social information. The panel depicted here illustrates the 'full social information/large group' treatment in which players benefitted from five constant sources of information. By clicking onto an anonymized name, players could see the innovation record of the corresponding player. By clicking onto an item (for example, the carved log outlined in the middle), players could observe the underlying combination that resulted in this item (depicted at the bottom of the left panel). Players from the partial information treatment did not benefit from the information depicted in the bottom of the left panel. Players from the small group treatment benefitted only from two constant sources of information. Players from the low connectivity treatment benefitted only from one changing source of information (among five). Isolated players could only observe their own record.

Source: https://doi.org/10.1038/ncomms9398 with permission (CC BY 4.0).

9780192886491_Book.indb 250
/12_first_proofs/first_proofs/xml_for_typesetting
27-Mar-25 23:02:49
OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

on specific datasets. To test the effect of group size, the model was run on a dataset comprising data from the 'individual learning', 'full social information/small group', and 'full social information/large group' treatments. 'Individuals' total score' was the dependant variable and 'group size' (1, 3, or 6) was introduced as a continuous independent variable, with 'group identity' as a random effect. To test the effect of group connectedness, the model was run on a dataset comprising data from the 'full social information/large group' and 'full social information/partially connected group' treatments. 'Individuals' total score' was the dependant variable and 'full connectedness' (0 or 1) was introduced as a binary independent variable, with 'group identity' as a random effect.

14.3.2.4 Results

Results indicate that individuals who are part of groups can produce totems that are more complex than any isolated individual can produce during the same amount of time. Moreover, the analyses of the different treatments reveal that this group-level ability to produce complex solutions is maximized when individuals are provided with full social information and when they are part of large and partially connected groups.

14.4 Limitations of these methods

Cultural evolution experiments have proved powerful in studying how individuals learn, transmit, and modify cultural information. Yet, as any method, cultural evolution experiments are associated with limitations of which experimenters must be aware. We go through the main limitations of these methods below.

14.4.1 Results can be task-specific

It can be argued that for many cultural evolution experiments, the results are heavily dependent on the idiosyncrasies of the specific task used in the experiment. Studies investigating the role of various transmission mechanisms on cumulative culture, for instance, have yielded inconsistent results. Some studies found that providing participants with the specific details of how an outcome was achieved has a strong effect on the pace of cumulative cultural evolution (Derex, Godelle, et al. 2013; Wasielewski 2014; Derex and Boyd 2015), while others did not (Caldwell and Millen 2009; Caldwell et al. 2012). Similarly, studies that have found empirical support for the transmission enhancing effects of teaching and language in the context of tool making (Morgan et al. 2015) have not consistently replicated across tool types (Putt et al. 2014; Whiten 2015; Pargeter et al. 2023). Part of the explanation is that cultural traits vary in their complexity and the extent to which specific transmission mechanisms are helpful to learners has been shown to vary with tool complexity (Lucas et al. 2020). Nevertheless, this illustrates the

9780192886491_Book.indb 251 /12_first_proofs/first_pr[illegible]/xml_for_typesetting 27-Mar-25 23:02:49

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

difficulty of choosing a task that appropriately captures the features of real-life cultural traits and poses the question of the ecological validity of the experimental tasks that are used to study cumulative cultural evolution (Miton and Charbonneau 2018; Derex 2021).

14.4.2 Lack of tasks' ecological validity

Drastic compression of the 'evolutionary' timescale is convenient for experimenters but it often forces them to rely on tasks that are simple compared to the type of problems that individuals must solve in the real world. This lack of complexity has been pointed out before (Miton and Charbonneau 2018; Caldwell et al. 2019; Derex 2021) and authors have argued that for a task to truly capture the complexity of most human technology, it must be opaque enough that one individual cannot decipher how to reproduce a solution without observing the underlying production process (e.g. Derex, Godelle, et al. 2013). For instance, the production of sophisticated stone tools requires a considerable amount of otherwise unobservable skills, such that a naive observer would not know how to produce these via observation alone. The same is not necessarily the case for paper aeroplanes, or many other cultural evolution tasks.

Another limitation that has been pointed out recently is that many tasks focus on the marginal improvement of already existing solutions (how fast can a plane fly, how fast can a wheel spin). These so-called optimization tasks prevent experimenters from studying actual innovation events through which novel behaviours and/or tools are created (Derex 2021). Compared to optimization, innovation events tend to result in more complex solutions which might be harder to learn than the pre-existing solutions they have been built upon. This suggests that the process of optimization might rely on a different set of sociocognitive abilities than the process of innovation, which has been arguably more important in our evolutionary history.

14.4.3 Cognition remains in a black box

One of the most prominent criticisms of cultural evolution experiments remains that the evolution of human cognition, and the cognitive mechanisms required for learning and transmitting these complex technologies, remain in a 'black box' (Clarke and Heyes 2016; Heyes 2016, 2018; Singh et al. 2021). That is, due to the experiments' reliance on cultural transmission alone, these methods have not helped to elucidate the precise cognitive mechanisms necessary for imitating and transmitting skills in general. For example, the cognition required to identify what a successful tool looks like, or what a successful behaviour looks like are left unexplored. In most experiments discussed so far, some marker of 'success' of either the tool or model is provided. This is an unlikely scenario in 'the real world' when deciding whom to learn from, or which item to copy, is crucial. Being able to recognize or assess what 'success' looks like is a given assumption

9780192886491_Book.indb 252 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:49

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

in most experiments. To give a crude and oversimplified example, when choosing whose canoe, or which canoe, to copy, perhaps it depends on which one goes fastest, how many people can fit inside, how long it lasts. How much variation in canoe design or success determines copying one person's canoe design/technique over another? Indeed, even this crude relationship between tool and success is oversimplified and does not capture the fact that *some* underlying knowledge of a tool/behaviour is needed to be able to aptly assess who or what is most successful compared to another, even if total causal understanding of the entire system is not (Derex et al. 2019).

14.5 Future Directions for Examining the Relationship between Language and Cultural Evolution

Surprisingly, the relationship between different forms of communication and cumulative cultural evolution has remained relatively understudied in the cultural evolutionary literature and experiments have only begun to scratch the surface of the complex relationship between communication and cultural evolution. In this section, we highlight avenues for improvement in our understanding of the relationship between language and cultural evolution.

14.5.1 The co-evolution between language and cumulative culture

Morgan et al.'s experiment demonstrates the importance of language for transmitting complex stone tool-making techniques (Morgan et al. 2015). However, results from another experiment that compared the acquisition of stone tool making among learners who were taught using speech alone (unassisted by gesture), gesture alone, or 'full language' (gesture plus speech) indicate that individuals who were taught using speech alone performed poorly compared to individuals instructed through either gesture alone or 'full language' (Cataldo et al. 2018). This suggests that learners might derive limited benefits from language, in the absence of demonstration, because the complex actions involved in skills may be too difficult to put into words. However, a scenario in which humans were communicating solely with language and in the absence of any gesture, body language, or physical demonstration seems wholly unrealistic for understanding our evolutionary past. Indeed, in this and many cultural evolution experiments, language is often either entirely absent, or full-blown modern-day human language is permitted. These are two unrealistic comparisons given that we know that human language went through prolonged periods of verbal protolanguage, most likely in co-evolution with gestural protolanguages (Jackendoff 1999; Bickerton 2007; Fitch

9780192886491_Book.indb 253 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:49

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

2017). Indeed, it is commonly argued that complex tool making and protolanguage co-evolved (Fitch 2010, 2017; Ghirlanda et al. 2017; Kolodny and Edelman 2018). For cultural evolution experiments to gain more ground in understanding not only the evolution of language but the role of communication in transmitting complex information in general, they will need to incorporate findings from cognitive science and language evolution to modify their methods. For example, incorporating 'protolanguage' conditions as a realistic comparison between full-blown language and gesture-only conditions.

14.5.2 The use of 'communicative gadgets' to promote learning

It has recently been hypothesized that cognitive mechanisms such as analogy, scaffolded by protolanguage, that allow the compression and communication of the kind of information needed for transmitting complex tool-making skills (see Brand et al. 2021 for review) might have played a significant role in the advent of cumulative cultural evolution. For example, a commonly used analogy for transmitting the tying of a bowline knot includes describing the string end as a rabbit, a loop as a burrow, and the other string end as a tree, so that the complex ordering of the action-sequence can be compressed and transmitted as 'the rabbit comes out the burrow, goes around the tree, and back down the burrow'. This is clearly easier to remember, communicate, and transmit than describing the precise actions of your fingers, hands, and each section of string in precise language. The use of 'communicative gadgets' such as analogies, recipes, stories, rules, and general principles to ease the memory load of learning and communicating complex information sequences will need to be investigated in future experiments (Ghirlanda et al. 2017; Kolodny and Edelman 2018). Diving into the cognitive mechanisms behind our ability to compress and chunk complex information, supporting our ability to socially learn with such high fidelity, will also require better integration with findings in cognitive science (Brand et al. 2021).

14.5.3 The role of communication in transmitting complex information

Finally, exploring our full range of communicative strategies, including body language, eye-contact, tone of voice, choice of language, gesture, and exaggeration that teachers use to emphasize certain actions or important details will also be crucial for a full understanding of our high-fidelity transmission abilities (Singh et al. 2021). In typical cultural evolution experiments, participants are often presented with a single task and have no other choice but to perform that task. In more realistic settings, participants might decide to give up on complex tasks and pursue simpler tasks, which could result

9780192886491_Book.indb 254 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:49

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

in the disappearance of hard to learn traits. Yet, types of communication can be used to encourage learners and support the acquisition of skills that require a large amount of deliberate practice (Stout 2005), which in turn may affect the probability of adoption of hard to learn traits. Overtly intentional communication (and particularly language) also allows potential learners to query what they do not understand and allows experienced individuals to explain, justify, and instruct, as appropriate to the needs of the learner. This might be especially important for behaviours that do not have immediate benefits, as inexperienced learners may be more likely to ignore them in favour of options that have more apparent benefits (Singh et al. 2021).

14.6 Conclusion

Experimental methods provide powerful ways to address a wide range of questions concerning the production, transmission, and maintenance of cultural traits. In this chapter, we have laid out the main principles and possibilities of cumulative cultural evolution experiments. To successfully implement these methods, certain core criteria must be present, and careful attention must be paid to ensuring the choice of task is appropriate for the question being asked. How transmission is implemented, how participants interact, and whether the task is presented physically or via computer are all options that are worth exploring, in accordance with how suitable each is to the research question. Current methods provide many rich opportunities for exploring the evolution of human cumulative culture but both experimental tasks that better reflect the complexity of human technology and experimental settings that implement evolutionarily relevant communication mechanisms are warranted. Furthermore, the propensity for humans to transmit complex knowledge and skills with high fidelity is not only reliant on communication strategies but is also a fundamental aspect of how language itself evolves. Methods that better integrate the findings from the fields of language evolution, cognitive science, and cultural evolution will be necessary in acquiring a full picture of how human behaviour, cognition, and communication has evolved, and is evolving.

Acknowledgements

M. D. acknowledges IAST funding from the French National Research Agency (ANR) under grant no. ANR-17-EURE-0010 (Investissements d'Avenir programme) and ANR funding through the ANR-21-CE28-0019-01 OPTILEARN project.

References

Acerbi, A., Mesoudi, A., and Smolla, M. (2022). *Individual-based Models of Cultural Evolution: A Step-by-step Guide Using R*. London: Routledge.

9780192886491_Book.indb 255 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:49

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Atkisson, C., O'Brien, M. J., and Mesoudi, A. (2012). 'Adult learners in a novel environment use prestige-biased social learning', *Evolutionary Psychology* 10(3): 147470491201000309, https://doi.org/10.1177/147470491201000309.

Beppu, A. and Griffiths, T. L. (2009). 'Iterated learning and the cultural ratchet', *Proceedings of the 31st Annual Conference of the Cognitive Science Society*.

Bickerton, D. (2007). 'Language evolution: A brief guide for linguists', *Lingua* 117(3): 510–526.

Boyd, R. and Richerson, P. J. (1985). *Culture and the Evolutionary Process*. Chicago, IL: University of Chicago Press.

Boyd, R. and Richerson, P. J. (1996). 'Why culture is common, but cultural evolution is rare?', *Proceedings of the British Academy* 88: 77–93.

Boyd, R., Richerson, P. J., and Henrich, J. (2011). 'The cultural niche: Why social learning is essential for human adaptation', *Proceedings of the National Academy of Sciences of the United States of America* 108: 10918–10925, <ISI>://WOS:000292137100018.

Brand, C. O., Heap, S., Morgan, T. J. H., and Mesoudi, A. (2020). 'The emergence and adaptive use of prestige in an online social learning task', *Scientific Reports* 10(1): 12095, https://doi.org/10.1038/s41598-020-68982-4.

Brand, C. O., Mesoudi, A., and Smaldino, P. E. (2021). 'Analogy as a catalyst for cumulative cultural evolution', *Trends in Cognitive Sciences* 25(6): 450–461, https://doi.org/10.1016/j.tics.2021.03.002.

Caldwell, C. A. and Millen, A. E. (2008a). 'Experimental models for testing hypotheses about cumulative cultural evolution', *Evolution and Human Behavior* 29(3): 165–171, <ISI>://WOS:000255224600002.

Caldwell, C. A. and Millen, A. E. (2008b). 'Studying cumulative cultural evolution in the laboratory', *Philosophical Transactions of the Royal Society B-Biological Sciences* 363(1509): 3529–3539, <ISI>://WOS:000260581900006.

Caldwell, C. A. and Millen, A. E. (2009). 'Social learning mechanisms and cumulative cultural evolution: Is imitation necessary?', *Psychological Science* 20(12): 1478–1483, <ISI>://WOS:000272163600009.

Caldwell, C. A., Schillinger, K., Evans, C. L., and Hopper, L. M. (2012). 'End state copying by humans (Homo sapiens): Implications for a comparative perspective on cumulative culture', *Journal of Comparative Psychology* 126(2): 161–169, <ISI>://WOS:000304287600007.

Caldwell, C. A., Atkinson, M., and Renner, E. (2016). 'Experimental approaches to studying cumulative cultural evolution', *Current Directions in Psychological Science* 25(3): 191–195, https://doi.org/10.1177/0963721416641049.

Caldwell, C. A., Atkinson, M., Blakey, K. H., et al. (2019). 'Experimental assessment of capacities for cumulative culture: Review and evaluation of methods', *Wiley Interdisciplinary Reviews: Cognitive Science* 11(1): e1516, https://doi.org/10.1002/wcs.1516.

Cataldo, D. M., Migliano, A. B., and Vinicius, L. (2018). 'Speech, stone tool-making and the evolution of language', *Plos One* 13(1): e0191071, https://doi.org/10.1371/journal.pone.0191071.

Cavalli-Sforza, L. L. and Feldman, M. W. (1981). *Cultural Transmission and Evolution: A Quantitative Approach*. Princeton, NJ: Princeton University Press.

Chudek, M., Heller, S., Birch, S., and Henrich, J. (2012). 'Prestige-biased cultural learning: Bystander's differential attention to potential models influences children's learning', *Evolution and Human Behavior* 33(1): 46–56, https://doi.org/10.1016/j.evolhumbehav.2011.05.005.

Clarke, E. and Heyes, C. (2016). 'The swashbuckling anthropologist: Henrich on the secret of our success', *Biology & Philosophy* 1–17, https://doi.org/10.1007/s10539-016-9554-y.

9780192886491_Book.indb 256 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:49

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Dean, L. G., Vale, G. L., Laland, K. N., et al. (2013). 'Human cumulative culture: A comparative perspective', *Biological Reviews* 89: 284–301, https://doi.org/10.1111/brv.12053.

d'Errico, F. and Stringer, C. B. (2011). 'Evolution, revolution or saltation scenario for the emergence of modern cultures?', *Philosophical Transactions of the Royal Society B: Biological Sciences* 366(1567): 1060–1069, https://doi.org/10.1098/rstb.2010.0340.

Derex, M. (2021). 'Human cumulative culture and the exploitation of natural phenomena', *Philosophical Transactions of the Royal Society B: Biological Sciences* 377: 20200311, https://doi.org/https://doi.org/10.1098/rstb.2020.0311.

Derex, M. and Boyd, R. (2015). 'The foundations of the human cultural niche', *Nature Communications* 6: 8398, https://doi.org/https://doi.org/10.1038/ncomms9398.

Derex, M. and Boyd, R. (2016). 'Partial connectivity increases cultural accumulation within groups', *Proceedings of the National Academy of Sciences* 113(11): 2982–2987, https://doi.org/https://doi.org/10.1073/pnas.1518798113.

Derex, M., Beugin, M.-P., Godelle, B., and Raymond, M. (2013). 'Experimental evidence for the influence of group size on cultural complexity', *Nature* 503: 389–391, https://doi.org/https://doi.org/10.1038/nature12774.

Derex, M., Godelle, B., and Raymond, M. (2013). 'Social learners require process information to outperform individual learners', *Evolution* 67(3): 688–697, http://dx.doi.org/10.1111/j.1558-5646.2012.01804.x.

Derex, M., Bonnefon, J.-F., Boyd, R., and Mesoudi, A. (2019). 'Causal understanding is not necessary for the improvement of culturally evolving technology', *Nature Human Behaviour* 3(5): 446–452, https://doi.org/https://doi.org/10.1038/s41562-019-0567-9.

Fitch, W. T. (2010). *The Evolution of Language*. Cambridge: Cambridge University Press.

Fitch, W. T. (2017). 'Empirical approaches to the study of language evolution', *Psychonomic Bulletin & Review* 24: 3–33.

Ghirlanda, S., Lind, J., and Enquist, M. (2017). 'Memory for stimulus sequences: A divide between humans and other animals?' *Royal Society Open Science* 4(6): 161011.

Henrich, J. (2004). 'Demography and cultural evolution: How adaptive cultural processes can produce maladaptive losses—The Tasmanian case', *American Antiquity* 69(2): 197–214, <ISI>://WOS:000221068800001.

Henrich, J. (2015). *The Secret of our Success: How Culture is Driving Human Evolution, Domesticating our Species, and Making us Smarter*. Princeton, NJ: Princeton University Press.

Heyes, C. (2016). 'Blackboxing: Social learning strategies and cultural evolution', *Philosophical Transactions of the Royal Society B: Biological Sciences* 371(1693): 20150369, https://doi.org/doi:10.1098/rstb.2015.0369.

Heyes, C. (2018). 'Enquire within: Cultural evolution and cognitive science', *Philosophical Transactions of the Royal Society B: Biological Sciences* 373(1743): 20170051.

Jackendoff, R. (1999). 'Possible stages in the evolution of the language capacity', *Trends in Cognitive Sciences* 3(7): 272–279.

Kempe, M. and Mesoudi, A. (2014). 'An experimental demonstration of the effect of group size on cultural accumulation', *Evolution and Human Behavior* 35(4): 285–290, http://linkinghub.elsevier.com/retrieve/pii/S1090513814000294?showall=true.

Kolodny, O. and Edelman, S. (2018). 'The evolution of the capacity for language: The ecological context and adaptive value of a process of cognitive hijacking', *Philosophical Transactions of the Royal Society B: Biological Sciences* 373(1743): 20170052.

Lucas, A. J., Kings, M., Whittle, D., et al. (2020). 'The value of teaching increases with tool complexity in cumulative cultural evolution', *Proceedings of the Royal Society B: Biological Sciences* 287(1939): 20201885, https://doi.org/doi:10.1098/rspb.2020.1885.

9780192886491_Book.indb 257 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:49

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Mesoudi, A. (2011). 'An experimental comparison of human social learning strategies: Payoff-biased social learning is adaptive but underused', *Evolution and Human Behavior* 32(5): 334–342, <ISI>://WOS:000294105300005.

Mesoudi, A. and Thornton, A. (2018). 'What is cumulative cultural evolution?', *Proceedings of the Royal Society B: Biological Sciences* 285(1880): 20180712, https://doi.org/10.1098/rspb.2018.0712.

Mesoudi, A. and Whiten, A. (2008). 'The multiple roles of cultural transmission experiments in understanding human cultural evolution', *Philosophical Transactions of the Royal Society B-Biological Sciences* 363(1509): 3489–3501, https://doi.org/https://doi.org/10.1098/rstb.2008.0129.

Miton, H. and Charbonneau, M. (2018). 'Cumulative culture in the laboratory: Methodological and theoretical challenges', *Proceedings of the Royal Society B: Biological Sciences* 285(1879): 20180677, https://doi.org/10.1098/rspb.2018.0677.

Morgan, T. J. H., Uomini, N. T., Rendell, L. E., et al. (2015). 'Experimental evidence for the co-evolution of hominin tool-making teaching and language', *Nature Communications* 6: 6029, https://doi.org/10.1038/ncomms7029.

Muthukrishna, M., Shulman, B. W., Vasilescu, V., and Henrich, J. (2014). 'Sociality influences cultural complexity', *Proceedings of the Royal Society B: Biological Sciences* 281(1774): 20132511, https://doi.org/10.1098/rspb.2013.2511.

Pargeter, J., Liu, C., Kilgore, M. B., et al. (2023). 'Testing the effect of learning conditions and individual motor/cognitive differences on knapping skill acquisition', *Journal of Archaeological Method and Theory* 30(1): 127–171, https://doi.org/10.1007/s10816-022-09592-4.

Putt, S. S., Woods, A. D., and Franciscus, R. G. (2014). 'The role of verbal interaction during experimental bifacial stone tool manufacture', *Lithic Technology* 39(2): 96–112, https://doi.org/10.1179/0197726114Z.00000000036.

Riede, F. (2014). 'Eruptions and ruptures: A social network perspective on vulnerability and impact of the Laacher See eruption (c. 13,000 BP) on Late Glacial hunter-gatherers in northern Europe', *Archaeological Review from Cambridge* 29(1): 67–102.

Singh, M., Acerbi, A., Caldwell, C. A., et al. (2021). 'Beyond social learning', *Philosophical Transactions of the Royal Society B: Biological Sciences* 376(1828): 20200050, https://doi.org/doi:10.1098/rstb.2020.0050.

Stout, D. (2005). 'The social and cultural context of stone-knapping skill acquisition', in V. Roux and B. Bril (eds.), *Stone Knapping: The necessary conditions for a uniquely hominin behaviour*, 331–340. Cambridge: McDonald Institute for Archaeological Research.

Thompson, B., Opheusden, B. v., Sumers, T., and Griffiths, T. L. (2022). 'Complex cognitive algorithms preserved by selective social learning in experimental populations', *Science* 376(6588): 95–98, https://doi.org/doi:10.1126/science.abn0915.

Wasielewski, H. (2014). 'Imitation is necessary for cumulative cultural evolution in an unfamiliar, opaque task', *Human Nature* 25(1): 161–179, https://doi.org/10.1007/s12110-014-9192-5.

Whiten, A. (2015). 'Experimental studies illuminate the cultural transmission of percussive technologies in *Homo* and *Pan*', *Philosophical Transactions of the Royal Society B: Biological Sciences* 370(1682): 20140359, https://doi.org/doi:10.1098/rstb.2014.0359.

9780192886491_Book.indb 258 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:49

CHAPTER 15

AN INTEGRATIVE APPROACH TO EARLY SYMBOLIC EVOLUTION

Experimenting with the past

ISOBEL WISHER AND KRISTIAN TYLÉN†

15.1 Introduction

Over the last couple of decades, new kinds of controlled lab experiments have been utilized to investigate processes of language evolution (Nölle and Galantucci 2023). One of the advantages of experiments is that they can test the plausibility of causal hypotheses and thus provide a privileged window to the potential underlying mechanisms driving evolutionary processes inaccessible to cross-sectional and anthropological approaches. This kind of work has been pivotal in demonstrating how symbolic practices emerge and evolve to become systematic communication systems (and ultimately natural languages), without assuming major or sudden neurological changes (Nölle et al. 2018).

However, while the impact of this kind of experimental work has been mainly recognized in the fields of cognitive science and cultural evolution studies, the interest has been more limited in neighbouring fields of archaeology and anthropology. This is perhaps due to the poor ecological validity of the experimental paradigms, among other factors. The prototypical language evolution experiment often unfolds in impoverished lab environments using made-up toy tasks, for instance, of participants exchanging line drawings to communicate concepts to each other. In addition, recurrent findings—for instance that communication systems evolve linearly from an initial stage of reliance on iconic reference (Fay et al. 2014; Tamariz and Kirby 2016; Verhoef et al. 2016)—does not univocally correspond with current archaeological insights. Some of the

†The author, Kristian Tylen, has passed away.

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

earliest symbolic practices in our species appear to include abstract geometrical signs, such as the *c.*80,000-year-old engraved ochres at Blombos Cave (Henshilwood et al. 2009) with iconic signs appearing later and coexisting with abstract signs during the Upper Palaeolithic, *c.*45,000–15,000 years ago (Pettitt 2021).

Similarly, there have been attempts within archaeological research to propose interpretations or hypotheses regarding the origin and evolution of language capabilities, particularly within the 'cognitive archaeology' subdiscipline, which attempts to understand the cognitive mechanisms behind particular archaeological signatures (Renfrew 2012). Often, this research tends towards arguing for the presence of language in hominin species through implicit assumptions regarding other cognitive processes being precursors to linguistic abilities, e.g. working memory or theory of mind (Coolidge and Wynn 2005; Stade 2020) or arguing that certain behaviours would have required a fully functioning syntactical language, such as the caching of materials in a landscape, or the long-distance exchange of materials (Marwick 2003; Barham and Everett 2021). Due to the impoverished nature of the archaeological record, these kinds of interpretations often require several inferential 'leaps' (e.g. materials were cached in a landscape, caching materials requires encoded information to recover materials, encoding information in this way necessitates language) without necessarily having the appropriate theoretical or empirical grounding to support these inferences (Botha 2010). Consequently, it is rare that these discussions are explicitly placed within the context of previous language evolution studies or consider the full complexity of linguistic abilities. Approaches from cognitive science have also been more directly implemented within archaeological research to understand symbolically mediated behaviours, such as the most salient visual features of animal representations (Meyering et al. 2020) or using fMRI to understand whether patterns of brain activation for abstract patterns correspond to those for objects or words (Mellet et al. 2019). This interdisciplinary work has contributed significant insights into aspects of past symbolic behaviour but similarly faces challenges, particularly in appropriately implementing and interpreting results derived from methods in cognitive science (Hodgson 2020).

There is thus a need to bridge the epistemic gap between archaeological and experimental records and approaches, to better understand the processes and mechanisms that drove the evolution of symbolism. In this chapter, we discuss the method of using controlled experimental studies to understand the *cognitive affordances* of particular patterns, signs, or symbols, and how these kinds of approaches can be utilized to contribute deeper insights into the development of symbolic behaviour observed in the archaeological record. While oral language and gesture do not fossilize, the archaeological record documents other forms of symbolic behaviour consisting of portable and parietal marks. Often these materials are sparse in number and ambiguous with respect to their particular symbolic use (for example, whether they form part of an aesthetic, communicative, or ritual practice), and thus competing interpretations emerge without systematic and general agreed-upon criteria to weigh arguments and evidence (Botha 2008). However, by using methods designed to experimentally investigate the perceptual and cognitive implications of the structural properties of abstract marks, and how these change over time, we can derive new and more extensive amounts of

9780192886491_Book.indb 260 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:49

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

data to test hypotheses and support inferences about their symbolic functions and the adaptive pressures driving symbolic evolution (Currie 2018). These kinds of integrated approaches thus have the potential to address issues with fragmentary, limited evidence from the archaeological record—which would normally not facilitate insights into cognitive mechanisms, and provide tangible, archaeological evidence for the evolution of symbolic practices—usually inaccessible for language evolution studies.

15.2 Investigating the cognitive affordances of signs

The 'cognitive affordances' of signs is a new and developing concept. Building on Gibson's theory of affordances (e.g. how a shape can be perceived as 'graspable'; Gibson 1977), it pertains to the way properties of marks come to scaffold aspects of cognitive processing relevant for the semiotic function and particular context of use. Similar to the way manual tools are incrementally refined over time for better manipulation and functionality, we build on the assumption that symbolic forms undergo continuous refinements in adaptation to their function, as sign producers become more skilful in producing signs that effectively resonate with our cognitive system in function-specific ways (Tylén et al. 2020). For example, if one needs to caution someone about potential danger, the perceptual saliency of the sign is an essential property that helps grabbing and directing a person's attention. As the sign continues to be used in this context, it will undergo a process of continuous refinement to better enhance its perceptual saliency. However, from this idea also follows that different semiotic practices might engage different aspects of perception and cognition. When we observe signs to change aspects of their form over time, the particular direction of changes might be informative about aspects of the past use context and particular symbolic function of the signs. For example, if our sign warning about potential danger appears to change its properties over time to become more memorable—rather than perceptually salient—we might infer the context of its use has changed to one that requires the sign to be easily reproduced from memory. Thus, by investigating the cognitive affordances of a sign, we can support inferences and test hypotheses about its past contexts of use.

Here, we provide a series of experimental paradigms that can be used to study such affordances. More concretely, we will provide examples of experimental procedures that can be used to investigate the extent to which signs effectively engage with different aspects of cognitive processing in a way that enables and supports their role in relevant symbolic practices pertaining, for instance, to communication. Some of the experimental procedures have a big literature and are well attested with respect to their reliability and replicability. Others are rather new and have only been applied in a few or even a single study.

The experimental procedures can complement structural analyses/descriptions of signs and symbolic artefacts by measuring their impact on our cognitive systems, from

9780192886491_Book.indb 261 /12_first_proofs/first_pr/xml_for_typesetting 27-Mar-25 23:02:49

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

early processes of visual salience and attention to memory or the recognition of intentionality. The approaches make it possible to test concrete hypotheses regarding the underlying cognitive and social mechanisms shaping a novel sign system. While we focus on ways to experiment with the past in the sense of deriving experimental stimuli from archaeological and historical sources, the same experiments can also be used for sign systems collected as part of semiotic lab or transmission-chain experiments (cf. e.g. Fay et al. 2008). The experimental series provided here is not intended to be exhaustive but is likely to evolve as future studies identify additional affordances that require measurement. Also, we do not intend for the suite of experiment to be blindly applied on any case of evolving communication systems. The particular choice of experiments should be driven by conceptual motivation and hypotheses for the concrete case in focus.

Importantly, the application of experiments in this context rests on the assumption that data collected from contemporary human participants can—at least to some extent—inform inferences about the past perception, cognition, and behaviour of our Palaeolithic ancestors. This, of course, warrants caution. While there is no evidence that the human anatomical brain has changed significantly over the last 250,000 years, modern day human participants grow up in symbol-saturated worlds that are likely to functionally rewire their cognitive systems (Pinaud et al. 2006). In addition, even among contemporary humans, variation in perception and cognition have been suggested to be contingent on cultural and ecological factors (Henrich et al. 2010). Yet, a lot of the basic architecture and preferences of the human visual system appear to be of a more universal character shared with other primates (Hubel and Wiesel 1968; Orban et al. 2004; Bruner 2018), and there are many ongoing comparative efforts to explore similarities and differences between humans and primates that can potentially inform the credibility of the experimental approach proposed here (Matsuzawa 2009; Tomasello and Herrmann 2010).

15.2.1 The saliency of signs

One prediction we could present for a novel communication system is that signs evolve over time to become more salient to the prospective users in the sense of effectively engaging with their attentional systems. In other words, the effectiveness of a communicative sign is—at least to some degree—contingent on the extent to which it is easily recognized as a salient signal, distinguishable from background noise and non-symbolic behaviour (Brinck 2001). In the following, we will address aspects of visual saliency, but in principle, the same prediction would hold for signs relying on other sensory modalities.

Saliency can be defined as a measure of the relative 'importance' or 'value' of a stimulus for the perceiver (Treue 2003). A salient stimulus is thus one that has an immediate emotional response and/or attraction on low-level attentional mechanisms. Often saliency is related to evolutionary factors such as threat detection, food, or mating, but there is

9780192886491_Book.indb 262 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:49

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

also evidence that certain aesthetic properties or other forms of value of a stimulus can make it 'stand out' from the background (Ramachandran and Hirstein 1999).

There exist different experimental procedures for measuring saliency, and it is not always clear how these measures relate to each other or rely on the same mechanisms. Since they also rely on different kinds of setup, the choice of method will also sometimes depend on available equipment. Here we will briefly introduce four measures: (i) continuous flash suppression; (ii) preferential looking; (iii) pupillometry; and (iv) the dot probe task.

15.2.1.1 Continuous flash suppression

Continuous flash suppression is a psychophysics technique thought to measure preconscious stages of attention (Wilke et al. 2003; Tsuchiya and Koch 2004; Yang et al. 2007; Wang et al. 2012). By using a mirror stereoscope, different visual stimuli organized side-by-side on a screen can be administered to each eye of the participant. Because the stimuli differ, the more dominant stimulus will initially override the less dominant stimulus target due to the phenomenon of binocular rivalry, and the participant will only consciously see the dominant stimulus. However, after a variable time, the less dominant stimulus will gain ocular dominance and enter conscious perception. This *time to emerge* is considered indicative of the relative saliency of the stimulus with shorter emergence times being associated with higher salience. Operationally, the experiment is conducted by flashing a colourful high-frequency suppressor stimulus to the dominant eye, while fading in the target semiotic stimuli to the less dominant eye. Participants are asked to report when they see the target by pressing a key.

This technique has, for instance, been used to show that fearful faces are more salient than emotionally neutral faces (Yang et al. 2007), or that the Kanizsa triangle is more salient than a scrambled version of the same graphical elements (Wang et al. 2012).

15.2.1.2 Preferential looking

Another experimental method for assessing the relative saliency of signs relies on eye-tracking and measures preferential looking. The idea is to repeatedly present pairs of competing signs in each side of the participant's peripheral vision and then record the relative, immediate attraction to one of them measured as the first fixation. When all pairwise combinations of stimuli are presented, it is possible to test predictions concerning their relative attentional attraction by identifying, for each trial, the stimulus of the first fixation. The technique has mostly been used with infants/children (Wildt et al. 2019), clinical groups (Wang et al. 2018), and primates (Lewis and Krupenye 2022), but there is no principled reason why it could not be used with regular adult participants (e.g. Chakrabarti et al. 2017).

15.2.1.3 Pupillometrics

Besides the spatial coordinates of saccades and fixations, most eye-tracking systems also provide measures of pupil-size. Under controlled light conditions, pupillometric measures can be informative of the relative impact of a stimulus, with dilations

9780192886491_Book.indb 263 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:49

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

being an expression of enhanced processing and/or arousal (higher saliency), while constrictions indicate reduced processing (lower saliency; Mathôt 2018, Mathôt and Vilotijević 2022). For instance, this technique has been applied to measure responses to the ostensive qualities of signals, with dilation responses to direct eye-contact or the calling of your name (Kampe et al. 2003; Tylén et al. 2012). It should be noted that the pupillometric signal can, at times, be difficult to predict and interpret due to the many competing accounts of its role in higher-level cognition (Ferencova et al. 2021). For instance, there might be cases where the test stimuli might appear more aesthetically arousing while the control stimuli still require more processing (Elschner and Gaissmaier 2019).

15.2.1.4 The dot probe task

Last, saliency can be measured by means of the so-called *dot probe task*. Similar to preferential looking, the method measures the relative attentional attraction of two competing stimuli presented in peripheral vision (Yang et al. 2012; van Rooijen et al. 2017). An advantage of the method is that it does not require specialized equipment like a mirror stereoscope or an eye-tracker (although some implementations use a touchscreen). In the dot probe task, participants are initially presented with a dot in the middle or lower middle of the screen to centre their attention. Then two stimulus images appear on each side of the screen for a short interval, typically 500 ms, after which they disappear, and a dot appears at the position of one of the images. The task of the participant is to indicate as fast as possible the position of the dot (left or right) either by pressing directly on it, if a touchscreen is used, or by pressing a corresponding key on a keyboard. The finding is that the presentation of a salient stimulus on the same side of the screen as the appearance of the dot has a faciliatory effect leading to faster response times since attention is already drawn to this side of the screen and, in contrast, that a salient stimulus presented on the opposite side of the dot can cause a delay in response. It should be noted that the majority of existing studies has used it to study highly emotional stimuli, for instance related to face perception (Xu et al. 2016). It should be noted, however, that there has recently been some controversy about the reliability of this task (Schmukle 2005; Chapman et al. 2019).

15.2.2 The intentionality of signs

Many communicative practices rely on materials and means that we also can encounter in non-communicative contexts (Tylén et al. 2011). An obvious example is how movements of our bodies can have a purely instrumental purpose, for instance, when we grab and move a cup to free space on the table, but additionally, how we can recruit similar bodily movements when we perform the communicative act of offering someone a cup of coffee (Scott-Phillips et al. 2009; Tylén et al. 2012). In the latter case, we will often mark the communicative intention of the act by means of eye-contact, velocity, or more exaggerated movement patterns. It is suggested that humans are very

9780192886491_Book.indb 264 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:49

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

sensitive to such ostensive qualities of behaviour that signify that someone is intending to address you in a communicative way (Gergely et al. 2007).

When applied to visual signs, we again recognize that a trace or mark can be a result of a coincidental natural or instrumental process, or that it can be purposefully crafted to address an audience with an aesthetic or communicative intent. The recognition of an ostensive intention is thus an important prerequisite for many semiotic practices; to interpret the meaning of a sign requires that it is indeed recognized as a sign by differentiating itself from natural, coincidental, or instrumental marks (Tylén et al. 2009, 2012). Evolving a new communicative system is also about expressing and recognizing the intention that your signs are to be perceived in a particular way. This could lead to the prediction that signs within a system should develop to become increasingly recognized as purposefully produced with semiotic intent, as the sender gains skill in making such ostensive qualities.

15.2.2.1 The intentionality recognition task

It can be hard to derive a reliable index of an abstract quality like intentionality using Likert scales or other rating procedures. It is simply unclear to most participants what constitutes either end of such scales (Brown 2000). However, when given the task to indicate which of two stimuli is more likely to be purposefully produced with communicative intent, the task suddenly becomes easy. That is, an effective way to measure the relative perceived intentionality of a series of signs is to present all pairwise combinations of signs in a two-item forced-choice task. In each trial, the participant is presented with two stimuli side-by-side on a screen and asked to indicate by pressing a corresponding key which of the two is more likely to have been purposefully produced (Tylén et al. 2020). In each case, the participant's choice is a function of properties of the chosen stimulus relative to the non-chosen one, and data can be analysed with an outcome contest model that takes both into account to indicate the ranked order of signs with respect to their intentionality (Manner 2016).

15.2.3 The memorability of signs

Another prediction that we could motivate for some communicative practices is that signs will evolve to become easier to learn, remember, and reproduce from memory. There has been a lot of discussion about the role of social learning and memory processes in the evolution of human cognition (Donald 1991; Van Schaik and Burkart 2011; Wynn and Coolidge 2011). While it is likely that human capacities for learning and recollection have evolved during the Palaeolithic, so have human skills in producing signs that resonate with such processes (Donald 2017). Many practices of communication require us to carry a repertoire of signs in memory to prepare us for situations of either communicative production or reception. When we observe signs to change properties of their form over time, these changes are not unlikely to be an adaptation for the human memory system (Kirby et al. 2008). In the following, we thus introduce

9780192886491_Book.indb 265 /12_first_proofs/first_pr[illegible]/xml_for_typesetting 27-Mar-25 23:02:49

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

examples of approaches to experimentally assess the relative learnability and/or memorability of signs. The techniques can thus, for instance, be used to test if a communication system evolves over time to become more memorable. We start by considering ways to investigate memory processes related to the production of signs, before we address aspects of recognition/reception and understanding.

15.2.3.1 The delayed reproduction task

In the context of sign reproduction, one way to test memorability is by presenting participants with a sign for a short interval. After a brief retention, the participant is instructed to reproduce the sign from memory. Subsequently the reproduction fidelity can be calculated by comparing the stimulus sign with the participant's reproduction as using structural similarity or error scores (Tamariz and Kirby 2015; Tylén et al. 2020). It is consistently observed that even with quite short retention times of a few seconds, some forms are easier to recall and reproduce than others, possibly due to aspects of symmetry, complexity, and order.

15.2.3.2 The decoding task

While the delayed reproduction task is only concerned with memorability for forms, we can also be concerned with the extent to which sign-users are able to learn and recall the meaning of signs (Fay et al. 2008). In the decoding task, participants are presented with a series of signs. First, they are trained on the individual form–meaning pairings by passive, self-paced exposure. Then they proceed to a testing phase, where each sign is presented, and participants select the associated meaning from a list. This procedure is iterated until they reach a particular threshold of accuracy (e.g. 80 per cent). In a third phase of the task, participants perform a forced-choice reaction time task. Each sign is presented for 40 ms followed by a mask. Then, a candidate meaning is presented on the screen, and the participant has to indicate as fast as possible if the sign is a match or a mismatch with the meaning by pressing one of two buttons on the keyboard. The method can thus test if certain signs are easier to learn and faster to detect than other signs (Fay et al. 2008).

15.2.4 The distinctiveness and discriminability of signs

Different practices of sign use might come with different affordances for discriminability. In a system of, for instance, abstract aesthetic decorations, it might be desirable that individual signs bear a certain degree of resemblance effectively making them less discriminable. In such systems, individual signs are often not associated with a denotational meaning (i.e. each sign does not point to an individual referent) but are appreciated for their visual form qualities. However, the decorative practice might serve to mark the identity of a group or an individual, that is, it might have a certain style or express a cultural tradition that can be recognized as associated with a particular group (Sackett 1977). That is, to the extent that a semiotic practice is an expression of a cultural

9780192886491_Book.indb 266 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:49

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

tradition, we might predict individual signs to evolve to share certain morphological features that mark their origin and belonging to a group, which makes the signs become more similar within the decorative practice of the group but increasingly different from decorative practices of neighbouring groups (Tylén et al. 2020).

In contrast, when signs are produced with communicative, denotational intent, different things might happen. Here, there might be a pressure for signs to increase their discriminability in order not to confuse signs with different meanings (Morin 2018). Importantly, the relative discriminability of signs is also a matter of perceptual learning, and what might appear as very similar signs to the novice might be perfectly discriminable to the expert (cf. e.g. the letters 'b' and 'd'). Still, when we encounter semiotic practices, we might formulate predictions about whether the signs within a system are likely to evolve to become more or less discriminable over time. In the following, we present three experimental paradigms that can provide complementary information about the relative similarity or discriminability of signs.

15.2.4.1 The sorting task

One way to investigate the emergence of style in a semiotic practice is by asking participants to perform a sorting task (Markman et al. 1981; Chollet et al. 2011). Signs from different sources are presented to the participant who is instructed to divide them in equally sized bins based on their perceived similarity. Participants' sorting accuracy is thought to be indicative of relative recognizability of style among signs belonging to the same system relative to that of other systems.

15.2.4.2 The similarity judgement task

In another version of the same general idea, participants are presented with multiple trials of a two-item forced match-to-sample paradigm. A target sign is shown in the middle of the lower part of the screen while two competitor signs are presented side-by-side above the target. One of the competitors is from the same sign system, while the other is a foil from a different system. The task of the participant is to indicate if the target sign belongs together with one or the other competitor by pressing the corresponding key. Again, responses are analysed with respect to accuracy (Tylén et al. 2020).

15.2.4.3 The timed discriminability task

A final experiment in this series again relies on a two-item match-to-sample task to investigate the relative discriminability of signs. This time, the main outcome variable is, however, reaction time rather than accuracy. Like in the case of the Similarity Judgement Task, a target sign is shown in the lower middle part of the screen while two competitor signs are shown on each side in the middle of the screen. One of the signs is identical to the target and the participant's task is, as fast as possible, to indicate which one by pressing the corresponding key. The trial-by-trial reaction time of the participant will depend on the relative discriminability of the two competitors with faster reactions times when the competitors are more discriminable (Tylén et al. 2020).

9780192886491_Book.indb 267 /12_first_proofs/first_pr…/xml_for_typesetting 27-Mar-25 23:02:49

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

15.2.5 The semantic transparency of signs

Communicative signs can refer to their meanings in several ways. One relatively well-established way to classify types of reference is by reference to Charles Sander Peirce's sign types, *icon*, *index*, and *symbol* (Peirce 1931; Deacon 1998). Compared to symbols, icons and indexes are considered more 'motivated' in the sense that they evoke more direct connections with their referents, by resemblance and correlation, respectively. When presented with a novel sign, we are often better off guessing the meaning of icons and indexes even without prior exposure or training, while symbols, which are thought to rely on arbitrary conventions are more opaque with respect to their meaning. In the context of the language evolution, iconicity and indexicality are thus thought to play an important role in the grounding of novel communicative systems, as the establishment of a novel convention necessitates a communicative code to negotiate the convention (Harnad 1990; Fay et al. 2013). However, although Peirce's sign types are often treated as categorical (a sign is either an icon or a symbol), there are observations from experimental semiotic studies that suggest that symbols can evolve from icons through a process of compression and conventionalization (Garrod et al. 2007). The suggestion is that while communicative parties initially will rely on information in the signs themselves to access their meanings (e.g. their resemblance with the referent), through repeated interactions, the signs can become mere cues to previous episodes in a shared history of successful interactions. This effectively allows communicators to compress their articulate signs to very parsimonious, minimally schematic cues with the consequence that they lose much of their iconic qualities and appear as arbitrary symbols (at least to the casual bystander).

From these observations, we could predict that novel communication systems will initially go through a grounding phase relying on more motivated (i.e. semantically transparent) means of reference, while over time conventionalizing to become more arbitrary (Fay et al. 2008). However, there might also be conditions making this process more or less pronounced. For instance, it has been found that vocabulary acquired early in life appears to have a higher degree of sound symbolism (i.e. iconicity), than vocabulary acquired late (Monaghan et al. 2014). We can investigate the relative semantic transparency by presenting signs to naive participants and have them try to guess their meanings.

15.2.5.1 The semantic transparency task

In this task, participants are presented with a series of signs one-by-one and asked to indicate their meanings. In Fay et al. (2008), the meanings were represented as a finite list of twenty concepts, but in other studies, it might be relevant to invite free text responses. Notice that this experiment is only relevant to the extent that it can be established what the 'correct' meaning is, which might not apply to all historical materials.

9780192886491_Book.indb 268 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:49

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

15.3 Applications

In the following, we provide a case study operationalizing the concept of cognitive affordances to investigate how an actual collection of signs documented in the archaeological records evolved over time. By experimentally measuring the cognitive implications of changes happening to the signs over a prolonged period of time, we can derive large quantities of new data to complement archaeological analyses of the material properties of the artefacts and support inferences about their symbolic function.

15.3.1 Understanding the evolution of symbolic behaviour at Blombos cave and Diepkloof rock shelter

Material remains from the deep past have significant potential for providing deeper insights into how language and symbolic behaviour evolved. Some of the earliest examples of probable symbolic behaviour derive from the MSA record in South Africa and consist of abstract patterns, engraved on ochre from the site of Blombos cave (dating to *c.*100,000–72,000 BP; Henshilwood et al. 2009) and engraved on ostrich eggshells from the Diepkloof rock shelter (*c.*109,000–52,000 BP; Texier et al. 2013). Archaeological research has previously attempted to interpret these material remains in relation to discussions about the emergence of language or symbol use (d'Errico et al. 2003; Henshilwood and Marean 2003), but the paucity of the archaeological record constrains the interpretations that can be yielded from these materials and often makes more conceptual discussions related to language or symbolic evolution difficult to substantiate (Botha 2008). This has resulted in several competing accounts regarding the nature of the abstract marks from Blombos cave and the Diepkloof rock shelter, from the suggestion that the artefacts served an aesthetic function (Hodgson 2014), markers of personal (Parkington et al. 2005) or group identity (Texier et al. 2013), through to the engravings constituting arbitrary symbols in a communication system (Henshilwood et al. 2009; Mellet et al. 2019). Despite different kinds of archaeological analysis being utilized to support a particular interpretation, these arguments often make implicit assumptions regarding the extent to which the marks may reflect symbolism or linguistic systems (Botha 2010).

To inform these discussions, Tylén et al. (2020) operationalized the existing competing accounts as testable hypotheses and conducted a series of cognitive experiments using outlines of the engravings as stimuli. The approach was grounded in the assumption that different symbolic functions (in this case, aesthetics, identity marking, and communication) are associated with different cognitive affordances. For instance, for a sign to effectively serve an aesthetic function, it should be salient to the visual system, but individual signs need not be highly discriminable between or within

9780192886491_Book.indb 269 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:50

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

the two archaeological sites. If, however, the signs also functioned as markers of group identity, there would likely be pressure on the system to evolve aspects of style, making it easier to recognize if a sign from later periods originated from one site or another than it is for signs from earlier periods. Last, for a system of signs to effectively function as a full-blown denotational system with each sign referring to each meaning, there would likely be pressure for the signs to become more discriminable over time.

Through a series of five experiments, it was found that over the course of approximately 40,000 years, the Blombos and Diepkloof engravings increased in visual saliency and perceived intentionality. Later engravings were also recalled and reproduced with higher fidelity than earlier engravings, and participants were more accurate in recognizing if they originated from Blombos or Diepkloof, suggesting that they evolved aspects of style over time. However, there was no indication that the engravings became easier to discriminate, which could suggest that they were not used as denotational symbols. Based on these observations, the authors thus presented support for the interpretation that the Blombos and Diepkloof engravings could have served an aesthetic function and could have signified group identity, while there was less support for the idea that they served as full-blown, denotational symbols.

15.4 Conclusion

In preceding sections, we have outlined an experimental approach to investigate the cognitive affordances of signs, that is, the propensity of signs to engage relevant aspects of cognitive processing related to their function in varies kinds of semiotic practices. The series of experiments makes it possible to test concrete hypotheses pertaining to new and evolving sign systems. These could be the product of semiotic lab experiments, using artificial toy languages (silent gesture, drawings, or slide whistles, etc.), but another strength of the approach is how it lends itself to an integrative approach directly engaging historical and archaeological artefacts as stimuli. By experimentally measuring the cognitive affordances of past symbolic artefacts, we can (i) add a new cognitive level of description relevant for human symbolically mediated behaviour, (ii) derive sizable amounts of data for quantitative treatment, and (iii) test concrete hypotheses about the mechanisms underlying the development of novel semiotic practices. Importantly, the approach is not reductionistic. It is not intended to replace careful archaeological or anthropological analyses, but to complement by adding new layers of evidence to support our inferences about past behaviour.

Acknowledgements

This work is supported by the EU ERC consolidator project eSYMb, 101044626 awarded to the second author, Kristian Tylén.

9780192886491_Book.indb 270 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:50

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

References

Barham, L. and Everett, D. (2021). 'Semiotics and the origin of language in the lower Palaeolithic', *Journal of Archaeological Method and Theory* 28(2): 535–579, https://doi.org/10.1007/s10816-020-09480-9.

Botha, R. (2008). 'Prehistoric shell beads as a window on language evolution', *Language & Communication* 28(3): 197–212.

Botha, R. (2010). 'On the soundness of inferring modern language from symbolic behaviour', *Cambridge Archaeological Journal* 20(3): 345–356, https://doi.org/10.1017/S0959774310000454.

Brinck, I. (2001). 'Attention and the evolution of intentional communication', *Pragmatics & Cognition* 9(2): 259–277.

Brown, J. D. (2000). 'What issues affect Likert-scale questionnaire formats', *Shiken: JALT Testing & Evaluation SIG Newsletter* 4(1).

Bruner, E. (2018). 'Human paleoneurology and the evolution of the parietal cortex', *Brain, Behavior and Evolution* 91: 136–147.

Chakrabarti, B., Haffey, A., Canzano, L., et al. (2017). 'Individual differences in responsivity to social rewards: Insights from two eye-tracking tasks', *PLoS One* 12(10): e0185146.

Chapman, A., Devue, C., and Grimshaw, G. M. (2019). 'Fleeting reliability in the dot-probe task', *Psychological Research* 83(2): 308–320.

Chollet, S., Lelièvre, M., Abdi, H., and Valentin, D. (2011). 'Sort and beer: Everything you wanted to know about the sorting task but did not dare to ask', *Food Quality and Preference* 22(6): 507–520.

Coolidge, F. L. and Wynn, T. (2005). 'Working memory, its executive functions, and the emergence of modern thinking', *Cambridge Archaeological Journal* 15(1): 5–26, https://doi.org/10.1017/S0959774305000016.

Currie, A. (2018). *Rock, Bone, and Ruin: An Optimist's Guide to the Historical Sciences*. Cambridge, MA: MIT Press.

d'Errico, F., Henshilwood, C., Lawson, G., et al. (2003). 'Archaeological evidence for the emergence of language, symbolism, and music—an alternative multidisciplinary perspective', *Journal of World Prehistory* 17(1): 1–70.

Deacon, T. W. (1998). *The Symbolic Species: The Co-evolution of Language and the Brain*. New York: WW Norton & Company.

Donald, M. (1991). *Origins of the Modern Mind: Three Stages in the Evolution of Culture and Cognition* (2013609). Cambridge, MA: Harvard University Press.

Donald, M. (2017). 'Key cognitive preconditions for the evolution of language', *Psychonomic Bulletin & Review* 24: 204–208.

Elschner, S. G. and Gaissmaier, W. (2019). 'What the pupils tell us about aesthetics (or not)', *Journal of Eye Movement Research* 12(7): 82.

Fay, N., Garrod, S., and Roberts, L. (2008). 'The fitness and functionality of culturally evolved communication systems', *Philosophical Transactions of the Royal Society B: Biological Sciences* 363(1509): 3553–3561.

Fay, N., Arbib, M., and Garrod, S. (2013). 'How to bootstrap a human communication system', *Cognitive Science* 37(7): 1356–1367, https://doi.org/10.1111/cogs.12048.

Fay, N., Ellison, T. M., and Garrod, S. (2014). 'Iconicity: From sign to system in human communication and language', *Pragmatics & Cognition* 22(2): 244–263, https://doi.org/10.1075/pc.22.2.05fay.

9780192886491_Book.indb 271 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:02:50

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Ferencova, N., Visnovcova, Z., Olexova, L. B., and Tonhajzerova, I. (2021). 'Eye pupil—a window into central autonomic regulation via emotional/cognitive processing', *Physiological Research* 70(Suppl 4): S669.

Garrod, S., Fay, N., Lee, J., et al. (2007). 'Foundations of representation: Where might graphical symbol systems come from?', *Cognitive Science* 31(6): 961–987.

Gergely, G., Egyed, K., and Király, I. (2007). 'On pedagogy', *Developmental Science* 10(1): 139–146.

Gibson, J. (1977). 'The concept of affordances', in R. Shaw and J. Bransford (eds.), *Perceiving, Acting, and Knowing: Toward an Ecological Psychology*, 67–81. New York: John Wiley & Sons.

Harnad, S. (1990). 'The symbol grounding problem', *Physica D: Nonlinear Phenomena* 42(1–3): 335–346, https://doi.org/10.1016/0167-2789(90)90087-6.

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). 'The weirdest people in the world?', *Behavioral and Brain Sciences* 33(2–3): 61–83.

Henshilwood, C. S. and Marean, C. W. (2003). 'The origin of modern human behavior—Critique of the models and their test implications', *Current Anthropology* 44(5): 627–651, https://doi.org/10.1086/377665.

Henshilwood, C. S., d'Errico, F., and Watts, I. (2009). 'Engraved ochres from the middle stone age levels at Blombos Cave, South Africa', *Journal of Human Evolution* 57(1): 27–47.

Hodgson, D. (2014). 'Decoding the Blombos engravings, shell beads and Diepkloof ostrich eggshell patterns', *Cambridge Archaeological Journal* 24(1): 57–69.

Hodgson, D. (2020). *Neuroimaging Does Not Support the Representational Nature of the Earliest Human Engravings: Comment on Mellet et al. (2019)* [Preprint]. Cambridge Open Engage, https://doi.org/10.33774/coe-2020-orfg8.

Hubel, D. H. and Wiesel, T. N. (1968). 'Receptive fields and functional architecture of monkey striate cortex', *The Journal of Physiology* 195(1): 215–243.

Kampe, K. K., Frith, C. D., and Frith, U. (2003). ' "Hey John": Signals conveying communicative intention toward the self activate brain regions associated with "mentalizing", regardless of modality', *Journal of Neuroscience* 23(12): 5258–5263.

Kirby, S., Cornish, H., and Smith, K. (2008). 'Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language', *Proceedings of the National Academy of Sciences of the United States of America* 105(31): 10681–10686, https://doi.org/10.1073/pnas.0707835105.

Lewis, L. S. and Krupenye, C. (2022). 'Eye-tracking as a window into primate social cognition', *American Journal of Primatology* 84(10): e23393.

Manner, H. (2016). 'Modeling and forecasting the outcomes of NBA basketball games', *Journal of Quantitative Analysis in Sports* 12(1): 31–41.

Markman, E. M., Cox, B., and Machida, S. (1981). 'The standard object-sorting task as a measure of conceptual organization', *Developmental Psychology* 17(1): 115.

Marwick, B. (2003). 'Pleistocene exchange networks as evidence for the evolution of language', *Cambridge Archaeological Journal* 13(1): 67–81, https://doi.org/10.1017/S0959774303000040.

Mathôt, S. (2018). 'Pupillometry: Psychology, physiology, and function', *Journal of Cognition* 1(16): 1–23.

Mathôt, S. and Vilotijević, A. (2022). 'Methods in cognitive pupillometry: Design, preprocessing, and statistical analysis', *Behavior Research Methods* 55: 3055–3077.

Matsuzawa, T. (2009). *Primate Origins of Human Cognition and Behavior*. Tokyo: Springer Science & Business Media.

Mellet, E., Salagnon, M., Majkić, A., et al. (2019). 'Neuroimaging supports the representational nature of the earliest human engravings', *Royal Society Open Science* 6(7): 190086.

9780192886491_Book.indb 272 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:50

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Meyering, L.-E., Kentridge, R., and Pettitt, P. (2020). 'The visual psychology of European Upper Palaeolithic figurative art: Using bubbles to understand outline depictions', *World Archaeology* 52(2): 205–222, https://doi.org/10.1080/00438243.2020.1891964.

Monaghan, P., Shillcock, R. C., Christiansen, M. H., and Kirby, S. (2014). 'How arbitrary is language?', *Philosophical Transactions of the Royal Society B-Biological Sciences* 369(1651), https://doi.org/10.1098/rstb.2013.0299.

Morin, O. (2018). 'Spontaneous emergence of legibility in writing systems: The case of orientation anisotropy', *Cognitive Science* 42(2): 664–677.

Nölle, J. and Galantucci, B. (2023). 'Experimental semiotics: Past, present, and future', in A. M. Garcia and A. Ibánez (eds.), *The Routledge Handbook of Semiosis and the Brain*, 66–81. New York: Routledge.

Nölle, J., Staib, M., Fusaroli, R., and Tylén, K. (2018). 'The emergence of systematicity: How environmental and communicative factors shape a novel communication system', *Cognition* 181: 93–104.

Orban, G. A., Van Essen, D., and Vanduffel, W. (2004). 'Comparative mapping of higher visual areas in monkeys and humans', *Trends in Cognitive Sciences* 8(7): 315–324.

Parkington, J., Poggenpoel, C., Rigaud, J.-P., and Texier, P.-J. (2005). From tool to symbol: The behavioural context of intentionally marked ostrich eggshell from Diepkloof, Western Cape', in F. d'Errico, L. Backwell, and B. Malauzat (eds.), *From Tools to Symbols: From Early Hominids to Modern Humans*, 475–492. Johannesburg: Wits University Press.

Peirce, C. S. (1931). *Collected Papers of Charles Sanders Peirce* (5942276). Cambridge, MA: Harvard University Press.

Pettitt, P. (2021). 'The origins of human visual culture: A three-stage hypothesis from babble to concordancy to inclusivity', in S. Gaudzinski-Windheuser and O. Jöris (eds.), *The Beef Behind All Possible Pasts: The Tandem-Festschrift in Honour of Elaine Turner and Martin Street*, vol. 2, 229–248. Leibnz: Römisch-Germanisches Zentralmuseum.

Pinaud, R., Tremere, L. A., and De Weerd, P. (2006). *Plasticity in the Visual System: From Genes to Circuits*. New York: Springer Science & Business Media.

Ramachandran, V. S. and Hirstein, W. (1999). 'The science of art: A neurological theory of aesthetic experience', *Journal of Consciousness Studies* 6(6–7): 15–51.

Renfrew, C. (2012). 'Towards a cognitive archaeology: Material engagement and the early development of society', in I. Hodder (ed.), *Archaeological Theory Today*, 2nd ed., 124–145. Cambridge: Polity Press.

Sackett, J. A. (1977). 'The meaning of style in archaeology: A general model', *American Antiquity* 42(3): 369–380.

Schmukle, S. C. (2005). 'Unreliability of the dot probe task', *European Journal of Personality* 19(7): 595–605.

Scott-Phillips, T. C., Kirby, S., and Ritchie, G. R. (2009). 'Signalling signalhood and the emergence of communication', *Cognition* 113(2): 226–233.

Stade, C. M. (2020). 'Theory of mind as a proxy for Palaeolithic language ability: An alternative to the search for the earliest symbolic material culture', *Language Dynamics and Change* 10(1): 59–85, https://doi.org/10.1163/22105832-20201000.

Tamariz, M. and Kirby, S. (2015). 'Culture: Copying, compression, and conventionality', *Cognitive Science* 39(1): 171–183.

Tamariz, M. and Kirby, S. (2016). 'The cultural evolution of language', *Current Opinion in Psychology* 8: 37–43.

9780192886491_Book.indb 273 /12_first_proofs/first_pr[illegible]/xml_for_typesetting 27-Mar-25 23:02:50

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Texier, P.-J., Porraz, G., Parkington, J., et al. (2013). 'The context, form and significance of the MSA engraved ostrich eggshell collection from Diepkloof Rock Shelter, Western Cape, South Africa', *Journal of Archaeological Science* 40(9): 3412–3431.

Tomasello, M. and Herrmann, E. (2010). 'Ape and human cognition: What's the difference?', *Current Directions in Psychological Science* 19(1): 3–8.

Treue, S. (2003). 'Visual attention: The where, what, how and why of saliency', *Current Opinion in Neurobiology* 13(4): 428–432.

Tsuchiya, N. and Koch, C. (2004). 'Continuous flash suppression', *Journal of Vision* 4(8): 61.

Tylén, K., Wallentin, M., and Roepstorff, A. (2009). 'Say it with flowers! An fMRI study of object mediated communication', *Brain and Language* 108(3): 159–166.

Tylén, K., Bjørndahl, J. S., and Weed, E. (2011). 'Actualizing semiotic affordances in a material world', in S. J. Cowley (ed.), *Distributed Language*, 81–100. Amsterdam: Benjamins Current Topics.

Tylén, K., Allen, M., Hunter, B. K., and Roepstorff, A. (2012). 'Interaction vs. observation: Distinctive modes of social cognition in human brain and behavior? A combined fMRI and eye-tracking study', *Frontiers in Human Neuroscience* 6: 331.

Tylén, K., Fusaroli, R., Rojo, S., et al. (2020). 'The evolution of early symbolic behavior in Homo sapiens', *Proceedings of the National Academy of Sciences* 117(9): 4578–4584.

van Rooijen, R., Ploeger, A., and Kret, M. E. (2017). 'The dot-probe task to measure emotional attention: A suitable measure in comparative studies?' *Psychonomic Bulletin & Review* 24: 1686–1717.

Van Schaik, C. P. and Burkart, J. M. (2011). 'Social learning and evolution: The cultural intelligence hypothesis', *Philosophical Transactions of the Royal Society B: Biological Sciences* 366(1567): 1008–1016.

Verhoef, T., Kirby, S., and de Boer, B. (2016). 'Iconicity and the emergence of combinatorial structure in language', *Cognitive Science* 40(8): 1969–1994.

Wang, L., Weng, X., and He, S. (2012). 'Perceptual grouping without awareness: Superiority of Kanizsa triangle in breaking interocular suppression', *PLoS One* 7(6), https://doi.org/10.1371/journal.pone.0040106.

Wang, Q., Hu, Y., Shi, D., et al. (2018). 'Children with autism spectrum disorder prefer looking at repetitive movements in a preferential looking paradigm', *Journal of Autism and Developmental Disorders* 48: 2821–2831.

Wildt, E., Rohlfing, K. J., and Scharlau, I. (2019). 'The role of saliency in learning first words', *Frontiers in Psychology* 10: 1150.

Wilke, M., Logothetis, N. K., and Leopold, D. A. (2003). 'Generalized flash suppression of salient visual targets', *Neuron* 39(6): 1043–1052.

Wynn, T. and Coolidge, F. L. (2011). 'The implications of the working memory model for the evolution of modern cognition', *International Journal of Evolutionary Biology* 2011.

Xu, M., Li, Z., Diao, L., et al. (2016). 'Contextual valence and sociality jointly influence the early and later stages of neutral face processing' *Frontiers in Psychology* 7: 1258.

Yang, E., Zald, D. H., and Blake, R. (2007). 'Fearful expressions gain preferential access to awareness during continuous flash suppression', *Emotion* 7(4): 882.

Yang, Z., Jackson, T., Gao, X., and Chen, H. (2012). 'Identifying selective visual attention biases related to fear of pain by tracking eye movements within a dot-probe paradigm', *PAIN®* 153(8): 1742–1748.

9780192886491_Book.indb 274 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:50

PART II

SIMULATIONS

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

9780192886491_Book.indb 276 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:50

CHAPTER 16

THE BAYESIAN ITERATED LEARNING MODEL

VANESSA FERDINAND

16.1 Making learning biases explicit

Throughout this book, we have seen many diverse examples of cognitive biases shaping the evolution of language and communication systems. This chapter covers the Bayesian iterated learning model (BILM), which was developed in response to early agent-based simulations of language evolution. The aim of these simulations was to determine whether structured languages could emerge from the process of culturally transmitting languages over several generations of general-purpose learning algorithms. The answer was yes, but there was a sneaking suspicion that biases in the learning algorithms were driving the emergence of structure. Unfortunately, it was difficult to determine what these biases were doing, because each one of these simulations used bespoke learning algorithms with built-in biases that were not straightforward to quantify and compare. The BILM made biases explicit and quantifiable by using Bayesian learners with mathematically specifiable learning biases in the form of a prior probability distribution over hypotheses about the language being learned (Griffiths and Kalish 2005, 2007). This innovation allowed researchers in language evolution, and cultural evolution more broadly, to precisely track the effect of learning biases on the dynamics of cultural evolution and the resulting structure of languages and artefacts that emerge from cultural transmission processes (see Kirby 2017, for review). To start off, let's take a brief look at the scientific backdrop against which the BILM was developed.

In the early 2000s, researchers were debating the role that learning biases play in shaping the evolution and structure of language. This debate was largely in response to Chomsky's claim that Universal Grammar (UG) placed strong constraints on the structure of language (Chomsky 1965); so strong that one could presumably read humanity's language-specific learning biases off the universally shared features of all human languages. Several computational models of language evolution were made in order to

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

formalize the various verbal arguments in this debate, so that these arguments could be quantitatively and systematically compared. The earliest computational models of language evolution were agent-based models, which were implemented in computer programs (e.g. Hurford 1989; MacLennan 1990) and interacting robots (e.g. Steels 1998). Agent-based modelling approaches to language evolution treat language as a complex adaptive system that emerges from the local rules that agents obey when they receive, process, and send linguistic utterances to one another (see Smith 2014, for review). These local rules give rise to language with emergent properties such as 'compositionality' or a 'Zipfian distribution' over word lengths, which is the result of local interactions, but measured at a different scale. Agent-based modelling allows researchers to relate the low-level mechanisms of language learning and use to the high-level structures of language, bridging an explanatory gap in language evolution known as 'the problem of linkage' (Kirby 1999). Early computational models of language evolution addressed the explanatory gap in the emergence of language at three different scales: evolution, culture, and learning (see Kirby et al. 2007; Smith 2018, for review). Research into the first scale focuses on the biological transmission of languages, where the language a learner speaks is genetically determined. In this framework, certain language types confer higher fitness to their speakers and spread through the population because those speakers have more offspring (Niyogi and Berwick 1997a, 1997b; Komarova et al. 2001; Nowak et al. 2000, 2001, 2002; Komarova and Nowak 2003). The second scale focuses on the cultural transmission of languages, where languages are repeatedly learned and passed on through successive generations in a process of *iterated learning* (Kirby 2001; Kirby et al. 2014). Here, the learning biases of agents and the cultural transmission process jointly shape the structure of languages that emerge. The third scale focuses on the development of language between pairs or groups of individuals during repeated interaction and the resulting language is shaped by individuals biases, communicative norms, and common ground (Steels 1998, 2001; Galantucci 2005; Scott-Phillips et al. 2009).

Several agent-based models and simulations at the second scale, cultural evolution, produced intriguing alternatives to the Chomskyan perspective that learning biases placed strong constraints on the structure of language. For example, many of these models showed that *compositionality*, a universal feature of human languages, could emerge from the process of cultural transmission itself, without any need to posit specific biases for it in the language learning algorithm (Batali 1998, 2002; Hurford 2000; Kirby 2001, 2002; Brighton 2002; Kirby and Hurford 2002; Smith et al. 2003). Compositionality emerged when a *learning bottleneck* was imposed between generations, such that agents had to induce their infinitely productive grammars from a small subset of data. When no such bottleneck was imposed, and agents received data on every signal–meaning pair in the entire language, compositionality did not emerge. The fact that several models, all implemented with different custom-made learning algorithms, produced the same result was a win: it showed that this result might be a general property of language evolution through learning bottlenecks. However, every learning algorithm has its biases, and it was possible that the algorithms used in all of these simulations had a bias in common that was actually driving this result. For example, Brighton's model contained an explicit bias towards concise grammars, following the principle of minimum description length (MDL) and Occam's

9780192886491_Book.indb 278 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:50

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Razor (see Brighton and Kirby 2001). This is a common property of learning algorithms and could be present, either explicitly or implicitly, in all of the learning algorithms in these studies. Smith (2003) systematically compared the algorithms used and found that, indeed, they shared two biases: a bias towards one-to-one signal–meaning mappings and a bias towards exploiting regularities in the input data. This then raised the question as to how much of the outcome of iterated learning is determined by the cultural transmission process and how much is determined by the learners' biases.

To answer this question with precision, one needs a way to quantify the strength of the learning bias and the strength of its effect on the outcome of iterated learning. This is tricky because learning algorithms are normally implemented in several lines of code that perform several operations. In a sense, an algorithm's bias is an emergent property of these operations and how they interact, meaning the bias cannot be attributed to a specific line of code nor manipulated in a straightforward manner. The fields of psychology and cognitive science do have methods for probing blackboxed systems, like cognition, to characterize its biases. For example, most learning algorithms can be cast as Bayesian inference problems, which is a modelling framework that has an explicit, quantifiable bias parameter and, by the 2000s, fitting such models to human data had already proven to be a fruitful way for cognitive science to understand human learning biases (Anderson and Milson 1989; Anderson 1990; Oaksford and Chater 1998; Chater et al. 1999; Tenenbaum and Griffiths 2001; Tenenbaum et al. 2011). So, researchers in language evolution could either fit Bayesian models to their learning algorithms to find out what their biases are, or they could just use agents whose learning algorithm *is* Bayesian inference. Enter, the Bayesian iterated learning model (Griffiths and Kalish 2005, 2007).

In the next two sections, I present the Bayesian iterated learning model, describe its components, and walk through its analysis. I have stuck closely to the form of the BILM found in the original journal article (Griffiths and Kalish 2007) and associated conference paper (Griffiths and Kalish 2005). I also created some code in R that implements a basic BILM simulation and analysis, so that you can jump right in to playing with one yourself. In the third and fourth sections, we will return to this historical narrative and then dissolve into speculation about the future of BILMs.

16.2 Components of Bayesian iterated learning

A Bayesian iterated learning model is any iterated learning model where the agents' learning algorithm performs Bayesian inference. Bayesian inference is a type of inductive inference that allows us to infer which underlying process generated a set of data we observed (Bayes 1763). In the case of language learning, the underlying process or hypothesis (h) is a language and the data (d) is a set of utterances.

Figure 16.1 shows the relationship between data and hypotheses within an iterated learning framework (adapted from the iconic original figure in Griffiths and Kalish 2007). Each learner receives data from the previous generation, infers which language

9780192886491_Book.indb 279 /12_first_proofs/first_pr[illegible]/xml_for_typesetting 27-Mar-25 23:02:50

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

was most likely to have generated the data, and then produces new data from the inferred language. Here, data are generated by languages with probability $p(d \mid h)$ and the languages are inferred from the data with probability $p(h \mid d)$. $p(d \mid h)$ is known as the *data likelihood* and is dictated by the structure of the language. $p(h \mid d)$ is known as the *posterior probability* and is calculated by the learner using Bayes' rule:

$$p(h \mid d) = \frac{p(h)p(d \mid h)}{\sum_i p(h_i)p(d \mid h_i)} \tag{1}$$

At its core, Bayes' rule is an equation that shows the proper way to flip a conditional probability: it shows you how to convert $p(A \mid B)$ to $p(B \mid A)$. Naive human reasoners[1] typically assume that $p(A \mid B) = p(B \mid A)$ (Tversky and Kahneman 1974), and this error has come to be known as the prosecutor's fallacy (Thompson and Schumann 2017). For example, the probability that someone is innocent given the evidence is not equal to the probability of the evidence given that someone is innocent (which is usually low). Equating these two quantities incorrectly ignores the base rate, which is the probability of being innocent in the first place (which is usually high). Likewise, the probability a particular utterance is generated by a language, $p(d \mid h)$, is not equal to the probability of a language given that utterance, $p(h \mid d)$: we must include the base rate, $p(h)$, which is the prior probability of the hypothesis.

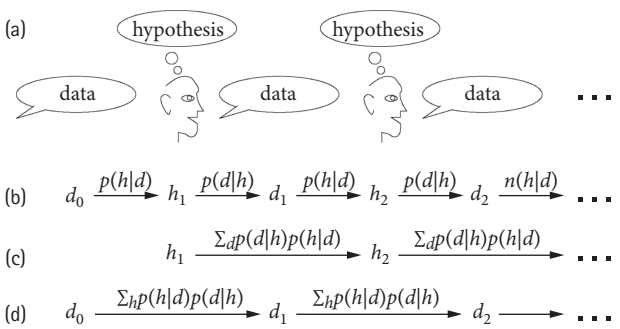


FIGURE 16.1 Schematic overview of Bayesian iterated learning.

a) Culture evolves as agents observe data, infer which hypothesis best describes the data, and generate new data from that hypothesis for the next learner. Data and hypotheses both evolve over time. b) Transitions from data to hypotheses ($d \rightarrow h$) are governed by the posterior probability $p(h|d)$ and transitions from hypotheses to data ($h \rightarrow d$ are governed by the data likelihood $p(d|h)$. c) Transitions between hypotheses can be calculated by summing over all intermediary data transitions and used to solve the stationary distribution over hypotheses. d) Transitions between data can be calculated by summing over all intermediary hypotheses transitions and used to solve the stationary distribution over data.

Source: Modified from Griffiths and Kalish (2007).

[1] Such as, everyone.

9780192886491_Book.indb 280 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:02:56

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

The *prior probability* represents the degree of belief that the learner assigns to each hypothesis before observing any data at all. The prior captures the learner's bias, making it both an *explicit* component of the learning algorithm and *quantifiable* by assigning each hypothesis a probability. For example, the relative strength of a bias towards one hypothesis over another can be quantified by comparing the relative prior probability of each hypothesis.

When $p(h \mid d)$ is computed for each of the hypotheses in the agent's space of possible hypotheses, then a *posterior distribution* is obtained, wherein each hypothesis is assigned a probability. The denominator in Bayes' Rule, $\Sigma_i p(h_i)p(d \mid h_i)$, simply normalizes the values in the posterior distribution to ensure that they all sum to one. It is important to note that the posterior distribution only contains information about the support that each hypothesis has relative to the other hypotheses being considered: if a new hypothesis were added to the mix, then the probabilities assigned to the other hypotheses would change.

Once the posterior distribution has been obtained, the Bayesian inference process is complete. Now the agent must do something with the posterior distribution. First, the agent uses the values in the posterior distribution to choose a language. Typically this hypothesis choice step is done by either *sampling* from the posterior distribution, by choosing a hypothesis in proportion to its posterior probability, or *maximizing* by choosing the hypothesis with the highest posterior probability. The original formulation of the Bayesian iterated learning model used the sampling strategy (Griffiths and Kalish 2005) and I discuss these two strategies further in section 16.4. Last, the agent generates data from the chosen hypothesis, governed by $p(d \mid h)$, to pass on to the next generation. This is generally done by sampling from $p(d \mid h)$.

In Figure 16.3, I have provided a brief code sample in R that implements Bayesian iterated learning. This simulation uses three types of hypotheses (h_1, h_2, h_3), three types of data (d_1, d_2, d_3), and arbitrary values for $p(h)$ and $p(d \mid h)$, which are shown in Figure 16.2. An initial data type is randomly chosen to serve as the input to the first learner, who then performs the following computations: 1) calculate the posterior using Bayes' rule; 2) choose a hypothesis type by sampling from the posterior; and 3) sample a data type from the chosen hypothesis. At the end of the generation, the chosen h and d are saved into respective timeseries variables, where the i^{th} element corresponds to the i^{th} generation, and the new d serves as the input to the next generation. This implementation corresponds to a linear transmission chain, where each 'generation' is comprised of a single learner and culture is inherited vertically, as opposed to horizontally or obliquely (Cavalli-Sforza and Feldman 1981: ch. 3).

9780192886491_Book.indb 281 /12_first_proofs/first_pr…/xml_for_typesetting 27-Mar-25 23:02:59

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

prior $p(h)$

h_1	h_2	h_3
.2	.3	.5

data likelihood $p(d|h)$

$A =$

	d_1	d_2	d_3
h_1	.8	.1	.1
h_2	.2	.6	.2
h_3	.3	.3	.4

posterior $p(h|d)$

$B =$

	h_1	h_2	h_3
d_1	.43	.16	.41
d_2	.06	.51	.43
d_3	.07	.21	.72

transition matrix $p(h_t, h_{t+1})$

$Q =$

	$h_{1_{t+1}}$	$h_{2_{t+1}}$	$h_{3_{t+1}}$
h_{1_t}	.36	.20	.44
h_{2_t}	.14	.38	.48
h_{3_t}	.18	.29	.53

FIGURE 16.2 The main components of Bayesian iterated learning in matrix representation: the prior distribution over hypotheses, $p(h_i)$; the data likelihood distributions per hypothesis type, $p(d_j|h_i)$; the posterior distribution over hypotheses per data type, $p(h_j|d_i)$; and the transition probabilities between all pairs of hypotheses between generations t and $t+1$.

The example values shown are those used in Figure 16.3 and assumes that hypotheses are chosen by sampling from the posterior. The values for the prior and likelihoods are chosen by the researcher, whereas the posterior and transition matrix are calculated from those values.

```
# specify p(h)
prior <- c(.2,.3,.5)

# specify p(d|h), row = h, column = d
likelihoods <- matrix(0,3,3)
likelihoods[1,] <- c(.8,.1,.1)  # h1 generates d1, d2, d3 with p(.8,.1,.1)
likelihoods[2,] <- c(.2,.6,.2)  # h2
likelihoods[3,] <- c(.3,.3,.4)  # h3

# save resulting h and d per generation here
h_history <- c()
d_history <- c()

# kick the simulation off with a randomly chosen initial data type
d <- sample(1:3,1)

# Bayesian iterated learning loop
generations <- 10000
for (i in 1:generations) {
    # calculate p(h|d)
    posterior <- likelihoods[,d]*prior/sum(likelihoods[,d]*prior)

    # choose a hypothesis from the posterior
    h <- sample(1:3,1,prob=posterior) # sample
    #h <- which.max(posterior) # maximize (or use this strategy instead)

    # generate data from the hypothesis
    d <- sample(1:3,1,prob=likelihoods[h,])

    # save results
    h_history <- c(h_history,h)
    d_history <- c(d_history,d)
}

table(h_history)/sum(table(h_history))  # prevalence of each hypothesis
table(d_history)/sum(table(d_history))  # prevalence of each data type
```

FIGURE 16.3 Example code in R that implements Bayesian iterated learning for 10,000 generations using arbitrary values of $p(h)$ and $p(d|h)$.

9780192886491_Book.indb 282 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:03:00

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

16.3 Analysis of Bayesian iterated learning

Bayesian iterated learning models are typically analysed in one of two ways: numerically or analytically (see Figure 16.4). Any BILM can be assessed numerically (given enough time and compute power), but not all models can be solved analytically. As a rough guide, a BILM can be solved analytically when it conforms to a Markov chain (but see Griffiths and Kalish 2007, for discussion). In both cases, the goal is to assess the 'outcome' of iterated learning, which means understanding how often the system produces the various data types and chooses the various hypotheses.

For numerical analyses, the goal is to run the simulation a sufficient number of times to obtain a good estimate of what the system's behaviour is. In Figure 16.3, the complete history of chosen h are saved to a timeseries variable called h_history and likewise, the complete history of emitted d are saved to d_history. These variables can be converted to probability distributions, $\hat{q}(h)$ and $\hat{q}(d)$, showing the proportion of times each h or d appeared in the simulation. Running one simulation for a longer number of generations will provide a better estimate of these proportions. Running several simulations in tandem (as independent replications with the same parameter settings) will provide better estimates of the variability in the system's behaviour at each generation.

For analytical solutions, the goal is to calculate the exact proportions of h and d from the known quantities $p(h)$ and $p(d \mid h)$, without running any simulations at all. Let's refer to these exact proportions as $q(h)$ and $q(d)$, in contrast to the numerical estimates $\hat{q}(h)$ and $\hat{q}(d)$ described above. These calculations can be cumbersome to do by hand, but when the problem is put into a matrix representation, we can leverage the tools of linear algebra to make efficient calculations. Figure 16.2 shows the main components of Bayesian iterated learning in matrix format, populated with the values used in the example simulation in Figure 16.3.

To solve $q(h)$, we first must calculate the transition matrix among all pairs of $h_t \rightarrow h_{t+1}$, which tells us the probability of transitioning from each hypothesis to every other hypothesis in two successive generations. This probability is given by $p(h_t, h_{t+1}) = \Sigma_d p(d \mid h_t) p(h_{t+1} \mid d)$ and can be quickly solved by matrix multiplication. Figure 16.2 shows $p(d \mid h)$ and $p(h \mid d)$ in matrix format as matrices **A** and **B**, respectively. Here, each row in **B** is computed using Bayes' rule and shows the posterior distribution for each data type observed. Matrix multiplication of $\mathbf{A} \times \mathbf{B}$ implements $\Sigma_d p(d \mid h_t) p(h_{t+1} \mid d)$ for all pairs of hypotheses (h_t, h_{t+1}) and the resulting matrix is the transition matrix among hypotheses, **Q**, where each row is the hypothesis at generation t and each column is the hypothesis at generation $t+1$. For example, to solve $p(h_t = 1, h_{t+1} = 2)$, the first row of **A** is multiplied by the second column of **B** and the results are summed (over d): $(.8 \times .16) + (.1 \times .51) + (.1 \times .21) = .20$. Computing **Q** in this way assumes that agents are sampling from $p(d \mid h)$ when they generate data and sampling from $p(h \mid d)$ when they choose hypotheses.

9780192886491_Book.indb 283 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:03:09

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN


```
# a) numerically estimate the transition matrix among hypotheses
x <- table(h_history[-length(h_history)],h_history[-1])
x <- x/rowSums(x)  # t -> t+1 transitions (row is t, col is t+1)

# b) calculate p(h|d) matrix, row = posterior given one data type
posteriors <- matrix(0,3,3)
posteriors[1,] <- likelihoods[,1]*prior/sum(likelihoods[,1]*prior) # d1 seen
posteriors[2,] <- likelihoods[,2]*prior/sum(likelihoods[,2]*prior) # d2 seen
posteriors[3,] <- likelihoods[,3]*prior/sum(likelihoods[,3]*prior) # d3 seen

# c) analytically solve the transition matrix among hypotheses
Q <- likelihoods %*% posteriors

# d) solve the stationary distribution of Q with eigen decomposition
vec1 <- eigen(t(Q))$vectors[,1]
vec1 <- vec1/sum(vec1)
```

FIGURE 16.4 Example code in R showing basic numerical (a) and analytical (b)–(d) solutions to the Bayesian iterated learning model.

The transition matrix is an important object that contains a wealth of information about the dynamics and evolution of a stochastic process. Here, I will point out a few ways it can be used to understand the evolution of $q(h)$ and $q(d)$. If we represent the state of hypotheses at time t as a vector $\vec{q}_t(h_1, h_2, \ldots, h_n)$, then we can multiply this vector by $\mathbf{Q}$ to obtain the probability that the next agent will choose each hypothesis at time $t+1$, as in $\vec{q}_t\mathbf{Q} = \vec{q}_{t+1}$. If a single agent chooses h_1 at time t, then the state vector would be $\vec{q}_t = (1, 0, 0)$ and $\vec{q}_t\mathbf{Q} = (.36, .20, .44)$. However, this logic also extends to populations of iterated learning agents. If an infinite number of agents were in each generation, and each passed on one data point to a random agent in the next generation, $\vec{q}_t\left(h_1, h_2, h_3\right)$ would represent the frequency of each hypothesis type in the population. For example, if $\vec{q}_t = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$, then $\vec{q}_{t+1} = \vec{q}_t\mathbf{Q} = (.22, .29, .49)$, meaning if we started with the same number of each type of language in the population, then, in the next generation, 22 per cent of the population would be speaking language h_1, 29 per cent would be speaking h_2, and 49 per cent would be speaking h_3. We can also solve the population frequencies for an arbitrarily large number of generations into the future by using an n-step transition matrix. Multiplying $\mathbf{Q}$ by itself n times, yields transition probabilities of the form $p(h_t, h_{t+n})$. Therefore, $\vec{q}_t\mathbf{Q}^n = \vec{q}_{t+n}$ tells us what the population at time t will look like in n generations. For example, if $\vec{q}_t = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$ then $\vec{q}_{t+10} = \vec{q}_t\mathbf{Q}^{10} = (.2, .3, .5)$, meaning, in ten generations, 20 per cent of the population will be speaking h_1, 30 per cent will be speaking h_2, and 50 per cent will be speaking h_3. If you increase n to a ridiculously high number, you will notice that $\vec{q}_t$ does not change any further: the population has converged to this vector, which can be said to represent the 'outcome' the Bayesian iterated learning process. This vector is known as the *stationary distribution*: it is a special vector that returns itself when multiplied by the transition matrix. The stationary distribution is defined as

9780192886491_Book.indb 284 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:03:17

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

$\pi\mathbf{Q} = \pi$. In the case of our example, $\pi = (.2, .3, .5)$ and this vector should be familiar, because it is identical to the prior! This is the main result of the very first Bayesian iterated learning model (Griffiths and Kalish 2005) and is known as 'convergence to the prior'.

16.4 The 'convergence to the prior' result (and how to break it)

The first Bayesian iterated learning model provided a simple, straightforward, and elegant answer to the question of how learning biases shape languages. The answer: learning biases *fully* determine the outcome of cultural evolution. Because the stationary distribution converges to the prior, this suggested that 'information transmitted via iterated learning will ultimately come to mirror the minds of the learners' (Griffiths and Kalish 2007). This was a shocking result, to some, that came down in favour of the Chomskian view that learning biases place hard constraints on the structure of language.

This result immediately suggested that the iterated learning paradigm could be used as a tool to reveal the inductive biases of learners in any given cognitive domain. All that was needed was to replace the agents in the simulation with humans in the laboratory, run an experimental version of the iterated learning model, and participants' biases would gradually be revealed. Two studies were conducted to validate this method on well-known and established biases in human cognition: Kalish et al. (2007) showed that iterated learning could reveal well-known biases in function learning that favour positive linear relationships (e.g. Brehmer 1971, 1974; Busemeyer et al. 1997) and Griffiths et al. (2008a) showed it could reveal well-known biases in category learning and generalization (e.g. Shepard et al. 1961; Nosofsky et al. 1994; Feldman 2000). Further studies were conducted that used this method as a discovery procedure for revealing unknown biases: Mesoudi et al. (2006) showed that humans preferentially retain social information over non-social information, Ferdinand and Zuidema (2008a) replicated the iterated function learning experiment with participants who were students of logic and mathematics, revealing a different prior containing more complex functions, and Lewandowsky et al. (2009) used this method to gain further insight into the biases people use when making predictions about everyday events. It is worth noting here that this method has been in use since Bartlett's serial reproduction task (Bartlett 1932) for investigating memory from a social psychology perspective and was appreciated for its ability to reveal inductive biases, although that terminology was not used nor formulated in any Bayesian sense.

At the same time that the iterated learning experiment was developing as a bias discovery tool, Kirby et al. (2008) developed the first iterated learning experiment with human participants in the field of language evolution, which replicated the results of

9780192886491_Book.indb 285 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:03:18

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

the original agent-based models in this field and showed the emergence of compositional languages. This paradigm soon blossomed as a technique, not just for revealing linguistic biases, but for testing hypotheses about how these biases affect the evolution of language. It has provided insight into phonetics (Verhoef et al. 2014), phonology (Del Giudice 2012; Mailhot 2013), category learning (Silvey et al. 2019; Carr et al. 2020; Ferdinand and Perfors 2020), linguistic regularization (Reali and Griffiths 2009; Smith and Wonnacott 2010; Ferdinand et al. 2013, 2019; Saldana et al. 2021), signed languages (Motamedi et al. 2019), and silent gesture (Micklos 2016) in humans and has been used in animal behaviour as well (Fehér et al. 2009; Claidière et al. 2014). This is just a brief list of examples and we will leave this topic now because it diverges from our focus on Bayesian iterated learning models.

Back to the first Bayesian iterated learning model. The convergence to the prior result had a few mismatches with the behaviour of the first iterated learning simulations. First, the simulations had shown an effect of the learning bottleneck (i.e. how much data was passed between agents each generation) on the types of languages that emerged (i.e. holistic vs. compositional), but in the BILM, increasing the amount of data that was passed between agents only increased the speed with which the population converged to the prior; it did not affect the stationary distribution over language types in any way (Kirby et al. 2014). Second, the simulations had also shown that cultural evolution *amplified* the effect of weak biases, but the BILM showed a one-to-one correspondence between the prior bias on each language type and that type's prevalence in the population (Kirby et al. 2014). Were there any modifications that could be made to the BILM that could reconcile these mismatches?

New Bayesian iterated learning models were soon made with the express purpose of breaking the convergence to the prior result. The first modification was made to the hypothesis choice strategy. Instead of sampling from the posterior distribution (Griffiths and Kalish 2005), agents chose the maximum *a-posteriori* (MAP) hypothesis from the posterior distribution; Griffiths and Kalish 2007, Kirby et al. 2007). The languages of MAP agents did not converge to the prior and exhibited behaviour that reconciled the two mismatches mentioned above: MAP learners converged to a stationary distribution that amplified the biases in the prior and was also affected by the amount of information passed between agents (Griffiths and Kalish 2007; Kirby et al. 2007).

The convergence to the prior result was also broken by putting agents into populations where they could learn from multiple teachers (Ferdinand and Zuidema 2009; Smith 2009) instead of using single teachers in vertical transmission chains (Griffiths and Kalish 2005, 2007; Kirby et al. 2007). Ferdinand and Zuidema (2009) found that samplers in populations converged to stationary distributions that amplified the prior, and that the degree to which the prior was amplified was also modulated by the population size and the structure of the data likelihoods. Smith (2009) found that samplers in populations did not converge to the prior either. Instead, one language comes to predominate the population, with the winning language determined by a combination of the prior and the initial proportions of the languages in the population. Both of these models, however, shared a new implementation quirk: agents could now

9780192886491_Book.indb 286 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:03:18

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

receive data generated from two different hypotheses at the same time. In this case, the agent is forced to choose, from their limited repertoire, whatever hypothesis happens to rack up the most likelihood from the combination of data points generated from these two hypotheses. This violates a condition for Bayesian rational analysis, required for the mathematical proof of convergence to the prior for samplers in Griffiths and Kalish (2005) because the agent is not entertaining all possible hypotheses that could have generated the data they received (i.e. data from two or more languages). Burkett and Griffiths (2010) responded to this issue by constructing two BILMs with new hypotheses spaces: one where agents learned about the proportions of words directly (without grouping them into languages) and one where agents learned the distributions over languages in the population, rather than individual languages. Both examples reinstated the assumptions required for Bayesian rational analysis and both resulted in convergence to the prior. At the same time, however, a third BILM was constructed in Dediu (2009) for highly complex populations that had spatial structure, overlapping generations, birth, death, and migration processes, and an explicitly modelled genetic bias for different linguistic features. In this case, both samplers and MAP agents showed convergence to the prior. The complexity of this model was far higher than other BILMs in use at the time, so the mechanisms that drove this result remain unexplained.

Populations were also modified to contain heterogeneous agents, where each agent could have a different prior. Ferdinand and Zuidema (2008b, 2009) was the first study to investigate heterogeneous priors in a BILM and found that two-agent populations with heterogeneous biases were able to converge on the same stationary distribution as the bottleneck widened. Two versions of the heterogenous priors model were investigated: one where all agents were samplers, and another where all agents were MAP learners. Samplers converged to a simple average of individuals' biases, whereas MAP learners converged to a distribution over languages that was not a simple average of the priors. Navarro et al. (2017, 2018) found similar dynamics for heterogenous populations and also investigated combinations of learners with weaker and stronger biases in the same population. They found that learners with stronger biases had a disproportionate effect on the stationary distribution.

Other studies modified the BILM to include external information about the world in which Bayesian agents were situated and found that this also changed BILM dynamics. In the BILM's original formulation, the agents' only learning target was the data that their peers produced, creating an insular learning problem that contained no ground truth. This makes sense for language evolution, where there is no objectively correct language to be learned beside what your speech community considers language to be. But this assumption does not capture other forms of cultural evolution as effectively. For example, in opinion dynamics, there might be some ground truth about the world to be learned (i.e. the world is not flat), where information about this topic can be learned socially (by receiving the opinions of others) and non-socially (by making natural observations or conducting experiments). Beppu and Griffiths (2009) was the first study to add ground truth to the BILM, creating agents that received a mixture of data from other agents and a fixed source in the world. They found that the stationary

9780192886491_Book.indb 287 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:03:18

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

distribution does not converge to the prior, and furthermore, that the prior has no effect on the stationary distribution whatsoever. Eventually, learners choose hypotheses as if they were only observing data from the world, which makes sense because the information being passed among agents is flexible, whereas the information from the world was fixed. Beppu and Griffiths (2009) also added teaching to the BILM: they gave agents access to one anothers' posterior distributions, in analogy to teachers communicating their degrees of belief in certain hypotheses to pupils, and found that the iterated learning chain becomes equivalent to Bayesian updating (a general model of learning, not cultural evolution) and the stationary distribution converges to the single hypothesis that best describes the data from the world. Overall, these findings are at odds with the received function of cultural transmission in social learning theory, where socially learned information guides learners to find better solutions to problems faster (Hoppitt and Laland 2013). Instead, the BILM model showed that cultural transmission and individual learning have identical outcomes. However, two later studies, which fit BILMs to cultural transmission experiments with human participants, show that both the prior and the world affect the outcome of iterated learning. Perfors and Navarro (2014) conducted an iterated language learning experiment where the meaning space contained salient discontinuities. They found that languages evolved to place category boundaries along the discontinuities present in the world, resulting in a distribution over languages that depends both on the learners' prior and the structure of the world. Thompson and Griffiths (2021) conducted an iterated learning experiment where participants learned to make arrowheads via a computer interface with the goal of finding the arrowhead design with the highest payoff (in terms of calories produced from a virtual hunting game). Researchers quantified participants' pre-existing biases about which arrowhead designs would perform best and set the experimentally determined optimal design to differ from this bias. They found that iterated learned converged to a suboptimal design that lay between the bias and the objectively best arrowhead. This finding also ran counter to social learning theory, but in a different way: it showed that learning biases can limit the cumulative improvement of culturally evolving artefacts.

After these initial assumption-probing forays, BILMs continued to develop in a variety of directions for understanding various aspects of the evolution of language and culture. Connections between the BILM were made with other models of evolution, which included proving an equivalence to Wright-Fisher drift (Reali and Griffiths 2010), contrasting the BILM to selection-based models in cultural evolution (Griffiths et al. 2008b), investigating co-evolutionary dynamics between data and hypotheses (Ferdinand 2018), and analysing the rate at which languages lose influence from their last common ancestor (Rafferty et al. 2014). Beta-binomial BILMs were formulated to investigate the evolution of linguistic regularity (Reali and Griffiths 2009) and were applied to domain-general regularization processes in cultural evolution (Ferdinand et al. 2019). BILMs were used to gain deeper insight into how weak biases become amplified (Thompson et al. 2016), how cultural evolution can shield culturally transmitted behaviour from natural selection (Smith and Kirby 2008; Thompson et al. 2016), and how complex processes prevent cultural universals from

9780192886491_Book.indb 288 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:03:18

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

arising from learnability biases (Rafferty et al. 2013). Further progress was also made in understanding the dynamics of populations of Bayesian agents across different social network structures (Whalen and Griffiths 2017), in dynamic social networks (Chazelle and Wang 2019), in cases where agents could or could not track the identity of different speakers in the population (Smith et al. 2017), and in age-structured populations that affected cumulative cultural evolution (Kirby and Tamariz 2022).

An overarching framework also emerged for explaining the general behaviour of iterated learning models, in which culture is shaped by two competing pressures: learnability and expressivity. In the case of language, the first pressure relates to the fact that languages need to be easy to learn if they are to be passed on to new individuals and survive across generations, whereas the second pressure relates to the number of distinct meanings the language can encode and how informative or communicative the language is. These pressures are also referred to as simplicity vs. informativeness (Carr et al. 2020), compression vs. communication (Kirby et al. 2015), and simplicity vs. expressivity (Smith et al. 2013). One extreme example of a language that does not balance this trade-off is a language that uses the same word for every meaning. Such a language would have maximal learnability, but minimal expressivity or communicative value. In fact, languages like these are obtained in iterated language learning experiments with human participants when there is no communication step to the experiment and participants' only task is to recall the words that they have learned (Kirby et al. 2008). However, when a pressure for learnability and expressivity are both present, cultural evolution discovers good data *compressions* that balance this trade-off. Compositionality is an example of a good (i.e. lossless) compression, which preserves the number of distinct meanings but optimizes the way those meanings are encoded into letters, and the one-word-for-everything language described above is a bad (i.e. lossy) compression because it simplifies language at the expense of decodability. BILMs have been used to explore this trade-off by using priors that explicitly encode biases towards learnability and/or expressivity to understand the evolution of compositional grammars (Kirby et al. 2015) and the evolution of semantic category systems (Carr et al. 2020). It is also worth noting a parallel approach to understanding compression in cultural evolution through the *information bottleneck principle* (Zaslavsky 2019). This approach was developed outside of the BILM framework, but utilizes Bayesian agents in a cultural-evolutionary context, and has been used to explain the evolution of colour categories (Zaslavsky et al. 2018, 2020) and the structure of other semantic category systems (Zaslavsky et al. 2019).

16.5 The future of Bayesian iterated learning

What does the future of Bayesian iterated learning hold? One of the biggest things that the BILM has going for it is that it applies beyond language evolution, to any form of

9780192886491_Book.indb 289 /12_first_proofs/first_pr[illegible]/xml_for_typesetting 27-Mar-25 23:03:18

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

cultural evolution that is inductive or inferential in nature. Bayesian inference is a computational[2] level of analysis that can be realized in a myriad of algorithms detailing how humans, animals, or AI systems process information. Much of the discourse around the BILM has been aimed at demonstrating that inductive biases do place constraints on cultural evolution and determining that these constraints can be weak, modulated by other factors, and yet still have strong effects on the types of cultural artefacts that evolve. This discourse was originally in response to Chomsky's claims about UG, the language acquisition device, and its purported origin as a random, one-off genetic mutation[3] and provided an alternative explanation, grounded in quantitative models, for how linguistic structures arise from learners' biases.

The next steps, in my opinion, could be using BILMs to dig deeper into understanding how various systems of cultural evolution work, by embedding Bayesian agents into increasingly realistic models to understand the *causal role* of inductive biases in shaping the evolution of complex systems, and by using them to make *predictions* about the outcome of real-world cultural evolution. Being able to quantify inductive biases by using Bayesian agents should make it possible to trace the precise effect of agents' biases as they percolate through a complex system. However, complex systems are (somewhat by definition) difficult to predict, because they are composed of low-level processes that give rise to system-level emergent properties, which can then feed back onto the low-level processes (Weisbuch 1991). However, tools from complexity science and carefully designed experimental manipulations in computational models can map out causal relationships in this space. One good example can be found in the recent thesis by Song (2022), where Bayesian agents with different biases towards cooperativeness play a public goods game and culturally evolve solutions to a tragedy of the commons problem. Another complex systems domain for the BILM is collective intelligence and distributed problem solving (Krafft et al. 2016). Recent work has shown how cognitively limited agents can still achieve rational inference at the population level through cultural evolution (Hardy et al. 2022). Because both learning and evolution can be understood as information processing systems (Donaldson-Matasci et al. 2010; Adami 2012; Suchow et al. 2017), future work could combine information theory with the BILM framework to understand how information processing in individuals gives rise to (or competes against) information processing at the levels of populations and long-term evolutionary processes.

Moving forward, it could be useful to reflect on what the prior in a BILM is capturing. Any natural cognitive system is riddled with biases as it perceives, processes, and produces information. But when fitting a model with a single bias parameter to behavioural data, all of these biases get lumped together and reflected in a single term. This is useful for summarizing the bias of a cognitive system, but loses fine-grained distinctions

[2] In the sense of Marr's three levels of analysis: computation, algorithm, and implementation (Marr 1982).

[3] Though who on earth that random individual would be able to talk to is beyond me. I guess we'd better assume that mutation occurred in a set of twins?

9780192886491_Book.indb 290 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:03:18

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

that could increase a model's predictive power. In actuality, the prior represents everything the agent brings to the inference task at hand besides the data itself. This means the prior represents a learner's innate biases, learned biases, previous domain-specific experience, any perceptual or production biases, and even their affective state at the moment of induction.

It could also be useful to loosen our fixation on the prior and reflect on what the other components of the BILM are doing. They might map on to other parts of cognition that also shape the evolution of culture. The structure of the data likelihoods are particularly interesting: they have been shown to affect the outcome of iterated learning (Ferdinand and Zuidema 2008b, 2009) and have also been used in Bayesian models of cognition to explain human judgements of representativeness (Tenenbaum et al. 2001; Abbott et al. 2012). But what I find most intriguing about the likelihoods is how they map on to the concept of *fitness* from theoretical evolutionary biology. Bayesian inference is formally equivalent to replicator dynamics (originally noted in Harper 2009, Shalizi 2009; and later noted in Ferdinand 2018, Krakauer 2011, Suchow et al. 2017). In this equivalence, each part of the equation for Bayes' rule maps on to each part of the replicator dynamics equation: each hypothesis is equivalent to a genotype, the prior represents the 'population' of hypotheses at time t, and the posterior represents the 'population' at time $t+1$. What determines the relative fitness of hypotheses, meaning what causes them to gain differential support in a learner's mind, is simply the likelihood of the data under each hypothesis: the data offers different ratios of support to each hypothesis and the probability of each hypothesis under the prior updates by exactly that ratio. When put into a Bayesian updating framework, where the agent observes new data points over several rounds of learning (and each posterior becomes the prior for the next round of learning), the learner converges on the hypothesis that is best supported by the data. This process follows identical dynamics to the way that a single genotype (the one with the highest fitness) takes over a population within the replicator dynamics framework. It is interesting that we can point to a part of Bayesian inference that tells us what, essentially, the 'fitness of ideas' is, yet this component does not feature in our computational analyses of cultural evolution.

Another card in the future of the BILM might be relaxing the framework to accommodate bounded rationality (Simon 1990). One way to do this could be maintaining *procedural rationality* (Simon 2001) by using Bayes' rule (as a rational procedure) to correctly compute a posterior from the prior and data likelihoods, but dispensing with the requirements that the Bayesian inference procedure should 1) entertain all possible hypotheses, or 2) use the veridical probabilities with which each hypothesis would generate each data type (i.e. the 'true' likelihoods). As we discussed earlier, the first BILMs that allowed agents to learn from multiple teachers in a population broke the assumption that agents were entertaining all possible hypotheses, and consequentially also broke the convergence to the prior result (Ferdinand and Zuidema 2009; Smith 2009). However, this result was restored in Burkett and Griffiths (2010) by restoring agents' ability to entertain all possible hypotheses. Is the entertainment of all possible hypotheses a realistic or desirable feature of these models? In a recent

9780192886491_Book.indb 291 /12_first_proofs/first_pr…/xml_for_typesetting 27-Mar-25 23:03:18

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

paper, Woensdregt et al. (2021) found that many BILMs exhibit a combinatorial explosion in the hypothesis space which makes the problem of language learning intractable as the size of the signal–meaning space scales up. This not only demands unrealistic levels of cognitive complexity from the agents, but also makes our models poor explanatory tools. It is reasonable to assume that agents entertain a finite set of hypotheses. This assumption also makes things interesting when reflecting on the isomorphism between Bayesian likelihoods and biological fitness. When modelling replicator dynamics, a subset of all possible genotypes compete against one another and one gradually wins out. In a different subset of genotypes, it might take more time for the winner to take over the population, or another genotype might beat it out altogether. This is because the competition matters. For example, if a Bayesian agent was learning the bias on a coin from observing several flips, and the true coin was double-headed, the posterior distribution would evolve differently if the agent were entertaining all possible coin biases than if it were entertaining just three biases: a double-headed coin, a double-tailed coin, and an unbiased coin. The posterior would still converge on the correct answer, but its trajectory in getting there would differ. As an interesting side note, the physics of coin flipping only allows the latter three possible biases (Gelman and Nolan 2002), so, in this case, entertaining a subset of hypotheses is even rationally justified.

The same issue comes up in heterogenous populations where agents do not share the same set of hypotheses. Say agent A is producing data from hypothesis a, but agent B is entertaining a modified version of that hypothesis, a'. One agent can still learn from another even if they do not contain the same set of hypotheses: if agent A produces data from hypothesis a, then agent B will converge on a'. When the 'true' hypothesis is not in the mix, the best approximating model wins out. In fact, the idea that the true model is not in the mix is very close to the core of Bayesian approaches to model selection and is where Bayesian statistics truly shines in comparison to frequentist approaches (Burnham and Anderson 1998).

One last issue with entertaining all possible hypotheses is that it tends to automatically impose the *no free lunch theorem* (Wolpert and Macready 1995, 1997). One way to visualize this theorem in the BILM context is this: if we required all hypotheses to produce a perfectly representative set of data (i.e. produce an infinitely large set of data from the $p(d_i \mid h)$ distribution) and then summed up how often each data type was produced, we would end up with a uniform distribution over data. All data types would be equally likely, and likewise, all hypotheses would be equally likely under that dataset. When a subset of hypotheses is chosen where the hypotheses, on average, *do* produce one type of data more than another, then this sets up a situation where hypotheses exhibit differential fitness. Ferdinand and Zuidema (2008b) found that using the latter type of hypothesis structure could break the convergence to the prior result in MAP learners, even with an unbiased prior. However, the explanation for this result in terms of differential fitness among hypotheses was missing because the isomorphism between Bayesian inference and replicator dynamics was not yet publicly known (Harper 2009).

There is also unexplored potential for applying the BILM to animal communication systems that are socially learned and culturally transmitted. First, one would

9780192886491_Book.indb 292 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:03:19

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

need to measure the biases involved in some aspect of a species' perception, learning, and/or production of communicative signals and then use these quantified biases to predict the evolution of the signal over time. For example, superb lyrebirds (*Menura novaehollandiae*) are excellent vocal mimics, culturally transmit their mimicry to other lyrebirds (Robinson and Curtis 1996), and exhibit quantifiable biases in their vocal mimicry abilities (Dalziell and Magrath 2012). Zann and Dunstan (2008) found that lyrebirds show consistent biases in their mimicry of the two-part call of the male eastern whipbird (*Psophodes olivaceus*), which is comprised of a long monotone note followed by an abrupt broadband 'whipcrack' spanning approximately 7 kHz in 200 milliseconds. Adult lyrebirds do not hold the monotone as long as whipbirds do, nor are they able to hit the full frequency range of the whipcrack. These biases are more pronounced in younger lyrebirds, further confirming that these deviations are the result of cognitive biases which require more data and practice to overcome. Between 1934 and 1949, twenty-two lyrebirds were translocated from the Australian mainland to the island of Tasmania, where they are not native (Fleay 1952). Several mainland species that lyrebirds mimic, including the whipbird, are also not found on Tasmania. This translocation event created a natural experiment where we can observe the evolution of the lyrebird's rendition of whipbird calls in the absence of whipbird models. This is equivalent to removing 'ground truth' data (about real whipbird calls) from a cultural transmission system. In BILMs, we know that including ground truth data creates artefacts that are intermediate in form between the agents' biases and the truth, and removing ground truth data creates artefacts that evolve into full alignment with agents' biases. Therefore, I would predict that modern-day renditions of the whipbird by Tasmanian lyrebirds (if such calls have still survived after ninety years of isolation from the mainland population) would exhibit shorter monotones and narrower whipcrack bandwidths than modern-day renditions of the whipbird by mainland lyrebirds.

Having said all that, is it possible that the era of Bayesian iterated learning modelling is coming to an end? Will the Bayesian fad in cultural evolution go the way of the Bayesian fad in cognitive science? I remember when I returned to Edinburgh from a research visit to Berkeley in 2013, I pulled up my shirt sleeve in lab meeting to reveal a Sharpie-markered 'tattoo' of Bayes' rule on my bicep. Everyone gasped in total belief, because this was the vibe of the times. Computational cognitive science has since shifted focus from Bayesian inference towards topics in deep learning and (dare I type this at risk of outdating this chapter before it even goes to print) Generative AI. Algorithmic bias is a huge topic in digital ethics and explainable AI and the BILM could be a powerful modelling framework for both revealing algorithmic biases and modelling what effect these biases have on the complex human–computer interaction systems in which they operate. For example, we could ask questions like, what will happen when GenAI algorithms are trained more and more on their own output? GenAI art is being posted all over the internet and is now fair game to be used as part of the training data for further GenAI applications. The BILM tells us that these images will ultimately come to mirror the minds of these learning algorithms. Currently, GenAI algorithms

9780192886491_Book.indb 293 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:03:19

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

are notoriously bad at accurately representing hands on humans and animals, by adding extra fingers and folding joints to impossible angles. If trained on these inputs, we would expect this error to be amplified and become even more diagnostic of GenAI art. Perhaps, over time, the internet will become swamped with content that has ceased to carry any information about the outside world, becomes patently useless or uninteresting to human users, and we all leave the internet in droves as its new end users, an army of bots, troll the internet filtering a detritus of bit strings for any shred of information about the real world[4]—like what restaurant in your neighbourhood *actually* serves the best *báhn mì*, or how many fingers a chimpanzee *actually* has—while we're all picnicking in the sunlight and making eye-contact again.

References

Abbott, J., Regier, T., and Griffiths, T. (2012). 'Predicting focal colors with a rational model of representativeness', *Proceedings of the Annual Meeting of the Cognitive Science Society* 34: 60–65.

Adami, C. (2012). 'The use of information theory in evolutionary biology', *Annals of the New York Academy of Sciences* 1256(1): 49–65.

Anderson, J. R. (1990). *The Adaptive Character of Thought*. New York: Psychology Press.

Anderson, J. R. and Milson, R. (1989). 'Human memory: An adaptive perspective', *Psychological Review* 96(4): 703.

Bartlett, F. C. (1932). *Remembering: A Study in Experimental and Social Psychology*. Cambridge: Cambridge University Press.

Batali, J. (1998). 'Computational simulations of the emergence of grammar', in J. R. Hurford, M. Studdert-Kennedy, and C. Knight (eds.), *Approaches to the Evolution of Language: Social and Cognitive Bases*, 405–426. Cambridge: Cambridge University Press.

Batali, J. (2002). 'The negotiation and acquisition of recursive grammars as a result of competition among exemplars', in Ted Briscoe (ed.), *Linguistic Evolution Through Language Acquisition: Formal and Computational Models*, 111–172. Cambridge: Cambridge University Press.

Bayes, T. (1763). 'An essay towards solving a problem in the doctrine of chances', *Philosophical Transactions of the Royal Society of London* 53: 370–418.

Beppu, A. and Griffiths, T. L. (2009). 'Iterated learning and the cultural ratchet', *Proceedings of the Annual Meeting of the Cognitive Science Society* 31: 2089–2094.

Brehmer, B. (1971). 'Subjects' ability to use functional rules', *Psychonomic Science* 24(6): 259–260.

Brehmer, B. (1974). 'Hypotheses about relations between scaled variables in the learning of probabilistic inference tasks', *Organizational Behavior and Human Performance* 11(1): 1–27.

Brighton, H. (2002). 'Compositional syntax from cultural transmission', *Artificial Life* 8(1): 25–54.

[4] Credit to Carl Bergstrom (personal communication) for this particular imagery of the information apocalypse.

9780192886491_Book.indb 294 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:03:19

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Brighton, H. and Kirby, S. (2001). 'The survival of the smallest: Stability conditions for the cultural evolution of compositional language', *European Conference on Artificial Life* 592–601.

Burkett, D. and Griffiths, T. L. (2010). 'Iterated learning of multiple languages from multiple teachers', *Proceedings of the 8th International Conference on the Evolution of Language*, 58–65.

Burnham, K. P. and Anderson, D. R. (1998). *Model Selection and Inference: A Practical Information-Theoretic Approach*. Heidelberg: Springer-Verlag.

Busemeyer, J. R., Byun, E., Delosh, E. L., and McDaniel, M. A. (1997). 'Learning functional relations based on experience with input-output pairs by humans and artificial neural networks', in J. R. Busemeyer, E. Byun, E. L. Delosh, and M. A. McDaniel (eds.), *Knowledge, Concepts and Categories*, 405–437. Cambridge, MA: MIT Press.

Carr, J. W., Smith, K., Culbertson, J., and Kirby, S. (2020). 'Simplicity and informativeness in semantic category systems', *Cognition* 202: 104289.

Cavalli-Sforza, L. L. and Feldman, M. W. (1981). *Cultural Transmission and Evolution: A Quantitative Approach*. Princeton, NJ: Princeton University Press.

Chater, N., Oaksford, M., Chater, N., and Oaksford, M. (1999). 'Ten years of the rational analysis of cognition', *Trends in Cognitive Sciences* 3(2): 57–65.

Chazelle, B. and Wang, C. (2019). 'Iterated learning in dynamic social networks', *Journal of Machine Learning Research* 20: 1–28.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Claidière, N., Smith, K., Kirby, S., and Fagot, J. (2014). 'Cultural evolution of systematically structured behaviour in a non-human primate', *Proceedings of the Royal Society B: Biological Sciences* 281(1797): 20141541.

Dalziell, A. H. and Magrath, R. D. (2012). 'Fooling the experts: Accurate vocal mimicry in the song of the superb lyrebird, menura novaehollandiae', *Animal Behaviour* 83(6): 1401–1410.

Dediu, D. (2009). 'Genetic biasing through cultural transmission: Do simple Bayesian models of language evolution generalise?', *Journal of Theoretical Biology* 259(3): 552–561.

Del Giudice, A. (2012). 'The emergence of duality of patterning through iterated learning: Precursors to phonology in a visual lexicon', *Language and Cognition* 4(4): 381–418.

Donaldson-Matasci, M. C., Bergstrom, C. T., and Lachmann, M. (2010). 'The fitness value of information', *Oikos* 119(2): 219–230.

Fehér, O., Wang, H., Saar, S., et al. (2009). 'De novo establishment of wild-type song culture in the zebra finch', *Nature* 459(7246): 564–568.

Feldman, J. (2000). 'Minimization of Boolean complexity in human concept learning', *Nature* 407(6804): 630–633.

Ferdinand, V. and Perfors, A. (2020). 'The evolution of category systems within and between learners', *Proceedings of the Annual Meeting of the Cognitive Science Society* 42: 648–654.

Ferdinand, V. (2018). 'The coevolution of data and hypotheses in Bayesian cultural evolution', *Proceedings of the 12th International Conference on the Evolution of Language*.

Ferdinand, V., Kirby, S., and Smith, K. (2019). 'The cognitive roots of regularization in language', *Cognition* 184: 53–68.

Ferdinand, V. and Zuidema, W. (2008a). 'An experiment in iterated function learning', *ILLC Reports, University of Amsterdam* (PP-2008-53).

Ferdinand, V. and Zuidema, W. (2008b). 'Language adapting to the brain: A study of a Bayesian iterated learning model', *ILLC Reports, University of Amsterdam* (PP-2008-54).

9780192886491_Book.indb 295 /12_first_proofs/first_pr[illegible]/xml_for_typesetting 27-Mar-25 23:03:19

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Ferdinand, V. and Zuidema, W. (2009). 'Thomas' theorem meets bayes' rule: A model of the iterated learning of language', *Proceedings of the Annual Meeting of the Cognitive Science Society* 31: 1786–1791.

Ferdinand, V., Thompson, B., Kirby, S., and Smith, K. (2013). 'Regularization behavior in a non-linguistic domain', *Proceedings of the Annual Meeting of the Cognitive Science Society* 35: 436–441.

Fleay, D. (1952). 'Transferring lyrebirds to tasmanian forests', *Victorian Naturalist* 69: 60–63.

Galantucci, B. (2005). 'An experimental study of the emergence of human communication systems', *Cognitive Science* 29(5): 737–767.

Gelman, A. and Nolan, D. (2002). 'You can load a die, but you can't bias a coin', *The American Statistician* 56(4): 308–311.

Griffiths, T. L. and Kalish, M. L. (2005). 'A Bayesian view of language evolution by iterated learning', *Proceedings of the Annual Meeting of the Cognitive Science Society* 27: 827–832.

Griffiths, T. L. and Kalish, M. L. (2007). 'Language evolution by iterated learning with Bayesian agents', *Cognitive Science* 31(3): 441–480.

Griffiths, T. L., Christian, B. R., and Kalish, M. L. (2008a). 'Using category structures to test iterated learning as a method for identifying inductive biases', *Cognitive Science* 32(1): 68–107.

Griffiths, T. L., Kalish, M. L., and Lewandowsky, S. (2008b). 'Theoretical and empirical evidence for the impact of inductive biases on cultural evolution', *Philosophical Transactions of the Royal Society B: Biological Sciences* 363(1509): 3503–3514.

Hardy, M. D., Krafft, P. M., Thompson, B., and Griffiths, T. L. (2022). 'Overcoming individual limitations through distributed computation: Rational information accumulation in multigenerational populations', *Topics in Cognitive Science* 14(3): 550–573.

Harper, M. (2009). 'The replicator equation as an inference dynamic', *arXiv preprint arXiv:0911.1763*.

Hoppitt, W. and Laland, K. N. (2013). *Social learning: An Introduction to Mechanisms, Methods, and Models*. Princeton, NJ: Princeton University Press.

Hurford, J. R. (1989). 'Biological evolution of the saussurean sign as a component of the language acquisition device', *Lingua* 77(2): 187–222.

Hurford, J. R. (2000). 'Social transmission favours linguistic generalisation', in C. Knight, M. Studdert-Kennedy, and J. R. Hurford, *The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form*, 324–352. Cambridge: Cambridge University Press.

Kalish, M. L., Griffiths, T. L., and Lewandowsky, S. (2007). 'Iterated learning: Intergenerational knowledge transmission reveals inductive biases', *Psychonomic Bulletin & Review* 14(2): 288–294.

Kirby, S. (1999). *Function, Selection, and Innateness: The Emergence of Language Universals*. Oxford: Oxford University Press.

Kirby, S. (2001). 'Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity', *IEEE Transactions on Evolutionary Computation* 5(2): 102–110.

Kirby, S. (2002). 'Learning, bottlenecks and the evolution of recursive syntax', in T. Briscoe (ed.), *Linguistic Evolution through Language Acquisition: Formal and Computational Models*, 173–204. Cambridge: Cambridge University Press.

Kirby, S. and Hurford, J. R. (2002). 'The emergence of linguistic structure: An overview of the iterated learning model', in A. Cangelosi and D. Parisi (eds.), *Simulating the Evolution of Language* 121–147. London: Springer-Verlag.

9780192886491_Book.indb 296 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:03:19

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Kirby, S. and Tamariz, M. (2022). 'Cumulative cultural evolution, population structure and the origin of combinatoriality in human language', *Philosophical Transactions of the Royal Society B* 377(1843): 20200319.

Kirby, S. (2017). 'Culture and biology in the origins of linguistic structure', *Psychonomic Bulletin & Review* 24: 118–137.

Kirby, S., Dowman, M., and Griffiths, T. L. (2007). 'Innateness and culture in the evolution of language', *Proceedings of the National Academy of Sciences* 104(12): 5241–5245.

Kirby, S., Cornish, H., and Smith, K. (2008). 'Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language', *Proceedings of the National Academy of Sciences* 105(31): 10681–10686.

Kirby, S., Griffiths, T., and Smith, K. (2014). 'Iterated learning and the evolution of language', *Current Opinion in Neurobiology* 28: 108–114.

Kirby, S., Tamariz, M., Cornish, H., and Smith, K. (2015). 'Compression and communication in the cultural evolution of linguistic structure', *Cognition* 141: 87–102.

Komarova, N. L. and Nowak, M. A. (2003). 'Language dynamics in finite populations', *Journal of Theoretical Biology* 221(3): 445–457.

Komarova, N. L., Niyogi, P., and Nowak, M. A. (2001). 'The evolutionary dynamics of grammar acquisition', *Journal of Theoretical Biology* 209(1): 43–59.

Krafft, P. M., Zheng, J., Pan, W., et al. (2016). 'Human collective intelligence as distributed Bayesian inference', *arXiv preprint arXiv:1608.01987*.

Krakauer, D. C. (2011). 'Darwinian demons, evolutionary complexity, and information maximization', *Chaos: An Interdisciplinary Journal of Nonlinear Science* 21(3): 037110.

Lewandowsky, S., Griffiths, T. L., and Kalish, M. L. (2009). 'The wisdom of individuals: Exploring people's knowledge about everyday events using iterated learning', *Cognitive Science* 33(6): 969–998.

MacLennan, B. (1990). 'Synthetic ethology: An approach to the study of communication', in C. G. Langton, C. Taylor, J. Doyne Farmer, and S. Rasmussen (eds.), *Artificial Life II: The Second Workshop on the Synthesis and Simulation of Living Systems*, 631–658. Redwood City, CA: Addison-Wesley Publishing Company.

Mailhot, F. (2013). 'Modeling the emergence of vowel harmony through iterated learning', in A. C. L. Yu (ed.), *Origins of Sound Change: Approaches to Phonologization*, 247–61. Oxford: Oxford University Press.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: W. H. Freeman.

Mesoudi, A., Whiten, A., and Dunbar, R. (2006). 'A bias for social information in human cultural transmission', *British Journal of Psychology* 97(3): 405–423.

Micklos, A. (2016). 'Interaction for facilitating conventionalization: Negotiating the silent gesture communication of noun-verb pairs', *Proceedings of the 11th International Conference on the Evolution of Language*.

Motamedi, Y., Schouwstra, M., Smith, K., et al. (2019). 'Evolving artificial sign languages in the lab: From improvised gesture to systematic sign', *Cognition* 192: 103964.

Navarro, D. J., Perfors, A., Kary, A., et al. (2017). 'When extremists win: On the behavior of iterated learning chains when priors are heterogeneous', *Proceedings of the Annual Meeting of the Cognitive Science Society* 39: 847–852.

Navarro, D. J., Perfors, A., Kary, A., et al. (2018). 'When extremists win: Cultural transmission via iterated learning when populations are heterogeneous', *Cognitive Science* 42(7): 2108–2149.

9780192886491_Book.indb 297 /12_first_proofs/first_pr…/xml_for_typesetting 27-Mar-25 23:03:19

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Niyogi, P. and Berwick, R. C. (1997a). 'Evolutionary consequences of language learning', *Linguistics and Philosophy* 20: 697–719.

Niyogi, P. and Berwick, R. C. (1997b). 'A dynamical systems model for language change', *Complex Systems* 11(3): 161–204.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., et al. (1994). 'Comparing modes of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961)', *Memory & Cognition* 22(3): 352–369.

Nowak, M. A., Plotkin, J. B., and Jansen, V. A. (2000). 'The evolution of syntactic communication *Nature* 404(6777): 495–498.

Nowak, M. A., Komarova, N. L., and Niyogi, P. (2001). 'Evolution of universal grammar', *Science* 291(5501): 114–118.

Nowak, M. A., Komarova, N. L., and Niyogi, P. (2002). 'Computational and evolutionary aspects of language', *Nature* 417(6889): 611–617.

Oaksford, M. and Chater, N. (1998). *Rational Models of Cognition*. Oxford: Oxford University Press.

Perfors, A. and Navarro, D. J. (2014). 'Language evolution can be shaped by the structure of the world', *Cognitive Science* 38(4): 775–793.

Rafferty, A. N., Griffiths, T. L., and Ettlinger, M. (2013). 'Greater learnability is not sufficient to produce cultural universals', *Cognition* 129(1): 70–87.

Rafferty, A. N., Griffiths, T. L., and Klein, D. (2014). 'Analyzing the rate at which languages lose the influence of a common ancestor', *Cognitive Science* 38(7): 1406–1431.

Reali, F. and Griffiths, T. L. (2009). 'The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning', *Cognition* 111(3): 317–328.

Reali, F. and Griffiths, T. L. (2010). 'Words as alleles: Connecting language evolution with Bayesian learners to models of genetic drift', *Proceedings of the Royal Society B: Biological Sciences* 277(1680): 429–436.

Robinson, F. N. and Curtis, H. S. (1996). 'The vocal displays of the lyrebirds (Menuridae)', *Emu* 96: 258–275.

Saldana, C., Smith, K., Kirby, S., and Culbertson, J. (2021). 'Is regularization uniform across linguistic levels? Comparing learning and production of unconditioned probabilistic variation in morphology and word order', *Language Learning and Development* 17(2): 158–188.

Scott-Phillips, T. C., Kirby, S., and Ritchie, G. R. (2009). 'Signalling signalhood and the emergence of communication', *Cognition* 113(2): 226–233.

Shalizi, C. R. (2009). 'Dynamics of Bayesian updating with dependent data and misspecified models', *Electronic Journal of Statistics* 3: 1039–1074.

Shepard, R. N., Hovland, C. I., and Jenkins, H. M. (1961). 'Learning and memorization of classifications', *Psychological Monographs: General and Applied* 75(13): 1.

Silvey, C., Kirby, S., and Smith, K. (2019). 'Communication increases category structure and alignment only when combined with cultural transmission', *Journal of Memory and Language* 109: 104051.

Simon, H. A. (1990). 'Bounded rationality', in J. Eatwell, M. Milgate, and P. Newman (eds.), *Utility and Probability*, 15–18. London: Palgrave Macmillan UK.

Simon, H. A. (2001). 'Rationality in society', in N. J. Smelser and P. B. Baltes (eds.), *International Encyclopedia of the Social Behavioral Sciences*, 12782–12786. Oxford: Pergamon.

Smith, A. D. (2014). 'Models of language evolution and change', *Wiley Interdisciplinary Reviews: Cognitive Science* 5(3): 281–293.

9780192886491_Book.indb 298 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:03:19

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Smith, K. (2003). 'Learning biases and language evolution', *Proceedings of the Workshop on Language Evolution and Computation, 15th European Summer School on Logic, Language and Information*.

Smith, K. (2009). 'Iterated learning in populations of Bayesian agents', *Proceedings of the Annual Meeting of the Cognitive Science Society* 31: 697–702.

Smith, K. (2018). 'How culture and biology interact to shape language and the language faculty', *Topics in Cognitive Science* 12(2): 690–712.

Smith, K. and Kirby, S. (2008). 'Cultural evolution: Implications for understanding the human language faculty and its evolution', *Philosophical Transactions of the Royal Society B: Biological Sciences* 363(1509): 3591–3603.

Smith, K. and Wonnacott, E. (2010). 'Eliminating unpredictable variation through iterated learning', *Cognition* 116(3): 444–449.

Smith, K., Kirby, S., and Brighton, H. (2003). 'Iterated learning: A framework for the emergence of language', *Artificial Life* 9(4): 371–386.

Smith, K., Tamariz, M., and Kirby, S. (2013). 'Linguistic structure is an evolutionary trade-off between simplicity and expressivity', *Proceedings of the Annual Meeting of the Cognitive Science Society* 35: 1348–1353.

Smith, K., Perfors, A., Feh'er, O., et al. (2017). 'Language learning, language use and the evolution of linguistic variation', *Philosophical Transactions of the Royal Society B: Biological Sciences* 372(1711): 20160051.

Song, Y. (2022). *Bias and Decision-Making in Social Dilemmas*. Master's thesis, University of Melbourne.

Steels, L. (1998). 'The origins of syntax in visually grounded robotic agents', *Artificial Intelligence* 103(1–2): 133–156.

Steels, L. (2001). 'Language games for autonomous robots', *IEEE Intelligent Systems* 16(5): 16–22.

Suchow, J. W., Bourgin, D. D., and Griffiths, T. L. (2017). 'Evolution in mind: Evolutionary dynamics, cognitive processes, and Bayesian inference', *Trends in Cognitive Sciences* 21(7): 522–530.

Tenenbaum, J. B. and Griffiths, T. L. (2001). 'Generalization, similarity, and Bayesian inference', *Behavioral and Brain Sciences* 24(4): 629–640.

Tenenbaum, J. B., Griffiths, T. L., et al. (2001). 'The rational basis of representativeness', *Proceedings of the Annual Meeting of the Cognitive Science Society 23*.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). 'How to grow a mind: Statistics, structure, and abstraction', *Science* 331(6022): 1279–1285.

Thompson, B. and Griffiths, T. L. (2021). 'Human biases limit cumulative innovation', *Proceedings of the Royal Society B* 288(1946): 20202752.

Thompson, B., Kirby, S., and Smith, K. (2016). 'Culture shapes the evolution of cognition', *Proceedings of the National Academy of Sciences* 113(16): 4530–4535.

Thompson, W. C. and Schumann, E. L. (2017). 'Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and the defense attorney's fallacy', in P. Roberts (ed.), *Expert Evidence and Scientific Proof in Criminal Trials*, 371–391. London, New York: Routledge.

Tversky, A. and Kahneman, D. (1974). 'Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty', *Science* 185(4157): 1124–1131.

9780192886491_Book.indb 299 /12_first_proofs/first_pr[illegible]/xml_for_typesetting 27-Mar-25 23:03:19

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Verhoef, T., Kirby, S., and De Boer, B. (2014). 'Emergence of combinatorial structure and economy through iterated learning with continuous acoustic signals', *Journal of Phonetics* 43: 57–68.

Weisbuch, G. (1991). *Complex Systems Dynamics*. Redwood City, CA: Addison-Wesley Publishing.

Whalen, A. and Griffiths, T. L. (2017). 'Adding population structure to models of language evolution by iterated learning', *Journal of Mathematical Psychology* 76: 1–6.

Woensdregt, M. S., Spike, M., de Haan, R., et al. (2021). 'Why is scaling up models of language evolution hard?', *Proceedings of the Annual Meeting of the Cognitive Science Society* 43: 209–215.

Wolpert, D. H. and Macready, W. G. (1995). *No Free Lunch Theorems for Search* (tech. rep.). Technical Report SFI-TR-95-02-010, Santa Fe Institute.

Wolpert, D. H. and Macready, W. G. (1997). 'No free lunch theorems for optimization', *IEEE Transactions on Evolutionary Computation* 1(1): 67–82.

Zann, R. and Dunstan, E. (2008). 'Mimetic song in superb lyrebirds: Species mimicked and mimetic accuracy in different populations and age classes', *Animal Behaviour* 76(3): 1043–1054.

Zaslavsky, N. (2019). *Information-Theoretic Principles in the Evolution of Semantic Systems*. PhD thesis, The Hebrew University of Jerusalem.

Zaslavsky, N., Kemp, C., Regier, T., and Tishby, N. (2018). 'Efficient compression in color naming and its evolution', *Proceedings of the National Academy of Sciences* 115(31): 7937–7942.

Zaslavsky, N., Regier, T., Tishby, N., and Kemp, C. (2019). 'Semantic categories of artifacts and animals reflect efficient coding', *arXiv preprint arXiv:1905.04562*.

Zaslavsky, N., Kemp, C., Tishby, N., and Regier, T. (2020). 'Communicative need in colour naming', *Cognitive Neuropsychology* 37(5–6): 312–324.

9780192886491_Book.indb 300 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:03:19

CHAPTER 17

COMMUNICATION GAMES

Modelling language evolution through dyadic agent interactions

XENIA OHMER AND CHRISTINE CUSKLEY

17.1 Introduction

Communication games provide a useful conceptual tool to reduce complex linguistic interaction into a more tractable problem. In this chapter, we outline how communication games have been used to model language evolution through dyadic agent interactions: repeated interactions between artificial agents give rise to effective communication systems between those agents. These simulations are not meant to be a replication of complex human language in interaction. Rather, they provide us with a controlled way of simulating the relative contribution of many different parameters to the evolution and emergence of a functioning multi-agent communication system. In particular, communication games allow us to look closely at the role played by myriad factors in establishing effective communication systems: ranging from individual cognitive biases and population-level dynamics to more localized phenomena such as feedback and context within specific interactions.

We accomplish this by focusing on two specific implementations that serve as the basis for many computational models of language evolution: the Naming Game (NG), which grew out of embodied artificial intelligence research (Steels 1995), and the Lewis Signalling Game (LSG), which has its roots in game theory (Lewis 1969).[1] These are by no means the only possible ways of formalizing and implementing agent-based communication games

[1] Note that Lewis himself was interested in coordination between agents rather than language evolution. We use the term LSG broadly for implementations of communication games referring to his framework. The *obverter model* was, for example, one of the earliest computational works to do so (Oliphant and Batali 1997).

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

that are relevant to language evolution, and we do not aim to give a comprehensive historical overview of the field (see e.g. Spike 2018: ch. 1, for a more detailed review; and van Eecke et al. 2022 for a perspective that integrates multi-agent reinforcement learning). However, these two approaches provide a useful illustration of convergent results, and a clear demonstration of the potential of the method for studying language evolution, particularly aimed at researchers who may be new to agent-based modelling.

Below, we outline the skeletal dynamics of agent-based communication games within an information-theoretic framework (section 17.1.1), followed by more detailed discussions of specific NG and LSG implementations (section 17.2), including Python code examples. To illustrate how NG and LSG tend to set a different focus while relying on the same skeletal dynamics, we describe how both types of games have been applied to shed light on the role of sensory biases in the evolution of colour terms. Finally, we turn to strategies for measurement and limitations of the approach (section 17.3), followed by a brief sketch of how these games have been used to examine other complex linguistic phenomena (section 17.4).

17.1.1 Skeletal dynamics

Communication games are a popular framework for simulating how agents learn or create a language. They are used to simulate grounded, interactive language and to examine the dynamics of emergence and change in large populations. In a communication game, there are at least $M \geq 2$ agents with either pre-specified signal–meaning mappings which adapt over time, or the capacity to store mappings as they encounter them. The game requires the agents to exchange signals, with the aim of developing signal–meaning mappings that overlap completely, or sufficiently for a receiver to reconstruct a sender's intended meaning given a particular signal (potentially situated in context, e.g. within a particular environment which may resolve signal ambiguity). The agents use feedback to improve communication over repeated interactions.

Here and throughout, we use information theoretic terminology to describe these games (Shannon 1948; Skyrms 2010). *Meaning(s)* are states or objects in a real or simulated external environment which are reflected in internal states of the sender or receiver. *Signal(s)* are the forms passed between agents, which map onto meanings—although mappings may not be shared between agents. *Signalling systems* are signal–meaning mappings, which vary in complex ways, e.g. involving context-specific interpretation, or many-to-one/one-to-many mappings. At this level, agent-based communication games share many properties with experimental communication games (see Raviv and Boeckx, Introduction, this volume) and broader work on artificial language learning (see Cuskley 2019). The list below illustrates the minimal elements of a dyadic agent-based communication game in terms of meanings and signals.

1. Sender agent S is tasked with communicating a particular target meaning, m_t.
2. S chooses a signal σ, to send to a receiver, R.

9780192886491_Book.indb 302 27-Mar-25 23:03:20

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

3. Given σ, and potentially other information (e.g. context), R tries to recover m_t, generating a guess m_r.
4. The outcome of the interaction is evaluated as a success or failure based on the relationship between m_t and m_r; in the simplest case, where $m_t == m_r$, the interaction is successful (and fails otherwise).
5. One or both agents update their internal signal–meaning mappings based on the outcome.

We introduce general notation and terminology here which we use throughout, but which may vary in some specific implementations. Below, we outline these dynamics in code (Algorithm 1), which connects to expanded code for specific implementations discussed in section 17.2.[2]

Algorithm 1 Skeletal dynamics

```
# Interaction structure for a single game
def interaction(S, R):
    m_t = getMeaning()
    sigma = S.getSignal(m_t)
    m_r = R.guessMeaning(sigma)
    if m_r == m_t:
        outcome = True
    else:
        outcome = False
    S.update(outcome, sigma, m_t, m_r)
    R.update(outcome, sigma, m_t, m_r)

# Run a full simulation with M agents for N interactions
def runSim(M, N):
    agents = initAgents(M)
    for _ in range(N):
        S, R = chooseAgents(agents)
        interaction(S, R)
```

The sender S is tasked with sending a signal such that R can reconstruct the sender's intended meaning. In some implementations, the receiver performs a specific action based on their interpretation. Based on the relationship between the sender's intended meaning and the receiver's interpretation of the signal, an interaction is minimally described as successful (m_t is recoverable given σ) or failed (m_t is not recoverable). As a result of the outcome of the interaction, one or both agents update their signalling systems; minimally, the receiver is given feedback about the outcome of the interaction, and updates their internal signal–meaning mappings accordingly.

There are key points at which we can introduce variation in these basic skeletal dynamics to investigate different elements of communication, and how these affect the resulting systems. First, while choosing S, R, and m_t are prerequisites for the interaction, *how* these are chosen can vary considerably. The choice of agents

[2] The code provided in this chapter can also be found in our repository, together with example simulations: https://github.com/XeniaOhmer/dyadic_agent_interactions/.

9780192886491_Book.indb 303 27-Mar-25 23:03:23

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

for each role, where there is one, is discussed in more detail below—but for some models, the choice of meaning is also important. Implementations may only involve a single meaning (see e.g. the Naming Game below), choose a meaning from a uniform distribution at each interaction (e.g. Franke 2016), or choose a meaning from, for example, a Zipfian distribution; Cuskley et al. 2018, Ohmer et al. 2020). Finally, *how S* chooses σ (GetSignal), *how R* interprets it (GuessMeaning), and how agents update their mappings, can involve complex parameterizations and learning algorithms. Variations in how these functions operate form a key difference for many implementations of these models.

In addition to this, we can distinguish games that involve only a single dyad from games played by populations of agents. The chooseAgents function can introduce variation: agents can be chosen randomly from the population for a role (Baronchelli 2016), or according to some pre-specified social network structure (Lu et al. 2009; Tria et al. 2015). In some implementations (including for dyads), sender and receiver roles are fixed, such that some agents can only play the role of *S or R* throughout the simulation (e.g. for basic LSG implementations as described below). In others, any agent may play either role over the course of a simulation, but is assigned the role of either sender or receiver during a specific interaction (Baronchelli 2016). Population-based games tend to focus on signalling systems in terms of the extent to which they are shared across agents, while single-dyad-based games are often framed in terms of understanding the dynamics of learning and internal states in individual agents. Below, we outline two basic implementations that differ in some of these dimensions, before discussing more complex applications of each (for additional applications, see Mudd and de Boer, Chapter 18, this volume).

17.2 Two specific implementations

17.2.1 Lewis Game

David Lewis (1969) introduced the framework of signalling games to explain how conventions of meaning can form in a society without an explicit agreement. Influenced by economist Thomas Schelling's work (1960) on cooperative games, he approached communication as a coordination problem between a sender and a receiver with shared interests. Within this framework, conventions are grounded in the repeated play of signalling games, where the sender and the receiver coordinate their strategies.

The LSG is a cooperative game, where the two agents have perfect common interest. The world can be in one of k states, and S can send one of $l \geq k$ signals to R to convey the current state. R, in turn, can perform one of k actions in response. Here, k corresponds to the number of possible values for m_t or m_r in the skeletal dynamics outlined above. In other words, the world state corresponds to the target meaning in the skeletal dynamics, and instead of guessing the target meaning itself, R has to guess the correct *action* for it. The correct action for each state is given by a pre-specified, agent-external

9780192886491_Book.indb 304 27-Mar-25 23:03:25

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

one-to-one mapping from states to actions, $f : \{m_t^1, \ldots, m_t^k\} \rightarrow \{m_r^1, \ldots, m_r^k\}$. These determine any feedback or updates. Although some variation is introduced by terminological conventions for the LSG (e.g. state, action), we adopt the notation from the skeletal dynamics described above for consistency. The game proceeds as follows:

1. A world state m_t is selected at random.
2. S (but not R) observes that state and sends a signal, σ, to R.
3. Based on σ, R performs an action m_r.
4. If m_r is the correct action given m_t, both agents receive positive payoff; otherwise payoff is zero.

A *successful* signalling system in the LSG is a communication strategy where the sender chooses a distinct signal for each state and the receiver performs the correct action for each signal according to f (the pre-defined, external action–state mappings).[3] Successful signalling systems form a Nash equilibrium, in other words, an evolutionarily stable state in which the agents cannot increase their payoffs by unilaterally changing their strategies. However, not every Nash equilibrium constitutes a successful signalling system. For example, in a pooling equilibrium, S always selects the same signal regardless of the state, and R always selects the same action regardless of the signal, which likely results in communicative failure.

Lewis himself arguably did not provide a formal explanation of how agents arrive at a signalling convention (i.e. no UPDATE function) and why they stick to it (Rescorla 2019). For a long time, (Lewis) signalling games have been analysed using *static* solution concepts from game theory, such as Nash equilibria and evolutionary stability (Huttegger et al. 2014). By now, however, they have been complemented with various game *dynamics*, including individual learning and evolution, such as reinforcement learning or the replicator–mutator dynamics (Skyrms 2010). Traditionally, learning-based approaches stay close to the original dyadic LSG and work with agents of varying cognitive complexity, while evolutionary approaches extend the game to populations and work with simpler agents (see e.g. Huttegger and Zollman 2011).

17.2.1.1 Python code

Algorithm 2 provides Python code for agent architectures for a simple LSG implementation with fixed roles and deterministic strategies. This interlocks with the skeletal dynamics described in section 17.1.1 as follows: CHOOSEAGENTS in Algorithm 1 return a persistent instance of SENDER/RECEIVER in Algorithm 2. To illustrate how different update functions can be plugged into the framework, we extended the basic LSG with a win-stay/lose-shift update rule (UPDATE).

[3] Note that Lewis used the generic term 'signalling system' for what we call 'successful signalling system' here. We maintain a distinction between the two to acknowledge that a stable signalling system, which is nonetheless not communicatively functional, can emerge under certain conditions, e.g. the pooling equilibrium (Huttegger et al. 2014).

9780192886491_Book.indb 305 27-Mar-25 23:03:27

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Following this update rule, agents keep their strategies if an interaction was successful and change their strategies if it was unsuccessful. When an unsuccessful interaction occurs, the sender randomly chooses a different signal for the given state, and the receiver randomly chooses a different action for the received signal. In other words, a failed interaction involves both agents altering their mappings. Win-stay/lose-shift is a simple example of various possible learning-based approaches and has been implemented as a baseline comparison for more complex learning (e.g. Barrett and Zollman 2009; Catteeuw et al. 2011; Barrett et al. 2017).

Algorithm 2 (Lewis) signalling game with k states, k actions, and l signals. The correct action for each state is predetermined as $f: m_i^t \rightarrow m_i^r$. In this example, the agents start with random strategies and we added a simple win-stay/lose-shift update mechanism.

```
import random

k =  # Define the number of states/actions.
l =  # Define the number of signals.

def getMeaning():
    return random.randint(0, k-1)

# Define Sender and Receiver class, including update functions.

class Sender:
    def __init__(self):
        self.strategy = {}
        for state in range(k):
            self.strategy[state] = random.randint(0, l-1)

    def getSignal(self, m_t):
        return self.strategy[m_t]

    def update(self, outcome, sigma, m_t, _):
        if not outcome:
            alternative_signals = list(range(l))
            alternative_signals.remove(sigma)
            self.strategy[m_t] = random.choice(alternative_signals)

class Receiver:
    def __init__(self):
        self.strategy = {}
        for signal in range(l):
            self.strategy[signal] = random.randint(0, k-1)

    def guessMeaning(self, sigma):
        return self.strategy[sigma]

    def update(self, outcome, sigma, _, m_r):
        if not outcome:
            alternative_actions = list(range(k))
            alternative_actions.remove(m_r)
            self.strategy[sigma] = random.choice(alternative_actions)
```

9780192886491_Book.indb 306 27-Mar-25 23:03:28

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

17.2.2 Naming Game

Lewis' interest in signalling games grew out of an interest in cooperation and coordination between agents; communication was the means by which coordination was achieved. A later approach, the Naming Game, focused more centrally on establishing successful communication between artificial agents, where coordination was the means by which communication was achieved. Unsurprisingly, this slight shift in conceptual focus led to some differences in implementation, but convergent results overall.

Seeking to look at potential cultural dynamics underlying language complexity, Steels (1995) began with what is now known as the Grounded Naming Game (Steels and Loetzsch 2012), using groups of spatially situated agents to examine how they built effective communicative conventions through repeated interaction. While this work began with 'populations' as opposed to the single-dyad approach initially used in LSGs, these populations were small, consisting of only five spatially situated agents, which could 'point' to other meanings or agents in the space. Each interaction is framed as a conversation consisting of:

1. *Initiation*: *S* chooses m_t;
2. *Communication*: *S* articulates a signal, σ;
3. *Reply*: *R* interprets σ and generates a guess m_r;
4. *Confirmation*: *S* evaluates m_r and agents update their internal mappings.

This aligns with the steps in the skeletal dynamics. These conversations occur repeatedly, each time with a random *S* and *R* chosen from the population. Where there is a mismatch between the sender's intended meaning and the receiver's guess, the sender has a means of 'pointing' to their intended target, engaging in a form of repair. Eventually, agents coordinate their vocabularies, converging on communicatively effective shared signalling systems. This groundedness has interesting relevance for language evolution, especially theories which foreground shared perceptual systems (Cuskley and Kirby 2013), repair (Dingemanse et al. 2015), and common ground (Scott-Phillips et al. 2009).

A simplified analytical approach, now more often known as the Naming Game, emerged from this work (see Baronchelli 2016 for an accessible summary). Like LSGs (complemented with learning), the 'ungrounded' NG involves agent updates triggered by a pre-defined threshold of communicative success (or failure) based on the match (or mismatch) of the meaning intended by *S* and the meaning guessed by *R*. The simplicity of this approach has led to a more expansive body of work, including clear analytical solutions for simple cases (Baronchelli 2016), and applications across fields like statistical physics (Castellano et al. 2009) and opinion dynamics (Verma et al. 2012). More complex implementations introduce variations on the number and nature of meanings (Lipowski and Lipowska 2009) or the network structure of agent populations (Lu et al. 2009), and even extensions into experiments with human participants (Centola and Baronchelli 2015). Analytical results show that convergence of a population on a random graph is highly predictable: the time to convergence on a shared vocabulary can be determined given the population size, the number of meanings, and the size of the signal space (Baronchelli et al. 2008).

9780192886491_Book.indb 307 27-Mar-25 23:03:29

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

17.2.2.1 Python code

Algorithm 3 provides Python code for a simple NG as outlined analytically in Baronchelli (2016). Key differences emerge to the LSG implementation in terms of how sender and receiver roles are defined, and in terms of how a meaning is 'chosen'. Here, chooseAgents selects two agents randomly from the population (with the additional constraint that *S* and *R* must be different agents). Because of this, only one Agent class is defined within the model, with separate functions for getSignal and guessMeaning. On the other hand, the choice of meaning for each interaction is deterministic: in this simple implementation, getMeaning returns only one value. This simplification works as long as we assume a bias against homonymy: where no two meanings can have the same label, all meanings are independent, allowing us to work with a single meaning without loss of generality (Baronchelli 2016). Thus, the game examines the formation of a single shared convention across a large population of agents (though note this has been extended to multiple meanings, see e.g. Lipowski and Lipowska 2009).

Algorithm 3 Basic NG simulation (with *l* signals). See Baronchelli (2016) for an accessible analytical description; see also Baronchelli et al. (2008).

```
import random

l =  # Define the number of signals.

def getMeaning():
    return 1

# Define single Agent class; any agent can play either S or R role at
# a given interaction.
class Agent:
    def __init__(self):
        self.signals = []

    def getSignal(self, _):
        if not self.signals: # Test if signals is empty.
            return random.randint(0, l-1)
        else:
            return random.choice(self.signals)

    def guessMeaning(self, sigma):
        if sigma in self.signals:
            return 1
        else:
            return 0

    def update(self, outcome, sigma, _1, _2):
        if outcome:
            self.signals = [sigma]
        else:
            self.signals.append(sigma)
```

17.2.3 Application: Sensory biases and category formation

Baronchelli et al. (2010) and Ohmer, Marino, et al. (2022) provide an example of how the NG (Baronchelli et al. 2010) and LSG (Ohmer, Marino, et al. 2022) frameworks can

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

be applied in different ways to understand similar linguistic processes. Both approaches study language-perception interactions, focusing on the relationship between perceptual biases and the formation of categories as expressed through emergent labels in a communication game. Here, we describe these briefly in terms of the elements identified in the skeletal dynamics, and the simpler implementations detailed above.

Baronchelli et al. (2010) use the NG to study the emergence of colour labels (see also Steels and Belpaeme 2005). Populations of $M = 50$ agents start without pre-defined colour categories, and a continuous colour space based on just-noticeable difference (JND) curves. Each agent has a dynamic inventory of signal–meaning associations, which initially comprises no signals and a single perceptual category encompassing all colours (i.e. a single meaning). As in the basic implementation of the NG, *S* and *R* are randomly sampled from the population. They face a scene of at least two colour stimuli, one of which is the defined value of m_t that *S* needs to communicate to *R*. If the agents cannot discriminate the target from the distractor(s), they refine (i.e. partition) their categorical repertoire and add new signals for new categories. Depending on whether the interaction is communicatively successful, signal–colour mappings are added to, or removed from, an agent's inventory.

These category and inventory updates follow pre-defined rules. If *S* cannot discriminate the target from the distractors given its current categorical repertoire, a new (discriminative) category is created for the target. New categories inherit all signals associated with the original category and acquire an additional new signal. If possible, *S* will send the signal for the target category which was last used successfully. Otherwise, they will send the new signal which was added to the category upon creation. For *R*, if the incoming signal is not in their inventory at all, the game fails. If *R* recognizes the signal, it randomly picks one of the categories associated with it. If this category includes the target (but not the distractors), the interaction is successful. Regardless of the outcome, *R* adds a new category if its current categories do not discriminate the target from the distractors. Updates occur as follows: in case of failure, *R* adds the received signal to the target's category. In case of success, other signals mapped to the target category are removed and only the successful signal is mapped to that category for both agents.

Using this setup, two classes of populations with differences in perception biases are compared. In 'human populations', the just noticeable difference (JND) function for colour stimuli corresponds to the perceptual biases of humans, whereas in 'neutral populations', perceptual biases are uniform across the colour space. The JND functions are encoded by only including targets and distractors that can be discriminated under the respective function in the same scene. Results show that a human JND function is sufficient for the emergence of well-documented typological patterns in colour naming (Berlin and Kay 1969), whereas a uniform JND function is not.

Ohmer, Marino, et al. (2022) employ the LSG framework to study the mutual influence between perceptual categories and emergent labels more generally. *S* and *R* are implemented as neural networks and trained with reinforcement learning. The two agents 'live' in a world of abstract three-dimensional (3D) objects, with different colours, sizes, and shapes, which are presented in the form of images. At each time step, *S* observes such an object, which constitutes the target for this interaction (m_t), and sends a signal (σ) to *R*. Based on the signal, *R* has to guess the target (m_r) from among a set of two or more distractors.

9780192886491_Book.indb 309 27-Mar-25 23:03:32

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Each agent consists of a neural network for processing images (vision module) and a neural network for generating/interpreting signals (language module). Both modules are called in Sender.getSignal (to map the target image to a signal) and Receiver.guessMeaning (to map the signal onto one of the target and distractor images), respectively (see Algorithm 2). The vision module is pre-trained on an object classification task while the language module is initialized randomly. Before running the experiments, agents with different perceptual biases are created by manipulating the pre-training procedure. As a result, agents in different conditions have different visual object representations. With respect to the skeletal dynamics (section 17.1.1), these differences highlight the fact that agents' internal states do not necessarily reflect meanings in the world in a static or predetermined way. In this particular case, biased agents are better at discriminating objects based on their colour, size, or shape.

The authors run several experiments, two of which will be highlighted here. In the first experiment, discrimination biases are fixed (i.e. the vision module parameters are frozen) and the effects of perceptual biases on emergent languages are observed. Results show that perceptual biases are translated into linguistic biases: if agents are good at discriminating objects along a specific dimension (e.g. colour), they prefer to communicate this dimension over others. In other words, emerging labels tend to group objects based on their similarity in this dimension. The second experiment, in turn, studies the influence of labels on category formation. Here, the agents can adapt their visual representations (which means the vision module parameters are trained) while learning to communicate. When learning biased languages, the agents' visual representations adapt to reflect linguistic categories. That is, objects with the same labels become perceived as more similar, and objects with different labels as more different. As a consequence, communicating agents align their visual representations. Similar co-adaptation effects between the formation of visual representations and language use have been observed in humans, such as grouping of visually similar objects under the same labels (e.g. Rosch et al. 1976), categorical perception effects where object representations adapt to linguistic categories (e.g. Goldstone and Herdrickson 2010), linguistic relativity more generally (Wolff and Holmes 2011), as well as the alignment of categories between communication partners on shorter timescales (Markman and Makin 1998; Suffill et al. 2019).

The results of these two studies are convergent and complementary. Both studies create experimental conditions where agents' perceived similarity of the same stimuli differs. Their results show how such differences in perception can lead to differences in categorization and emergent communication. Baronchelli et al. (2010) focus on colour perception and colour labels, which are well studied in humans and allow for a direct quantitative comparison. The NG is well suited to model population-level effects such as the emergence of colour naming universals. By implementing specific perceptual biases in populations of agents, they aim to investigate emergent patterns in linguistic systems. Ohmer, Marino, et al. (2022), in contrast, look at a set of visual properties, including colour, and study the mutual influence between perceptual similarity and linguistic categories in the representations of individual agents. The agents have integrated

9780192886491_Book.indb 310 27-Mar-25 23:03:32

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

perceptual and linguistic systems, which allows for an in-depth analysis of each agent's visual representations and the communication protocol. While looking at similar phenomena, the two studies are in line with the tradition of using the NG to investigate population-level phenomena using many simple agents, and the LSG to investigate phenomena at the level of the individual using smaller populations and agents with more internal structure.

17.3 Strategies for implementation

In this section, we provide some specific strategies for implementing agent-based communication games. First, we outline how to think about approaching measurement in a model, particularly how this may differ depending on whether you aim to quantify properties of learners, or properties of language systems. Second, we turn to the general strengths and weaknesses of the approach, which are especially useful to keep in mind for researchers new to the method.

17.3.1 Model measurements

Like any quantitative investigation, a key question in agent-based communication games is what will be measured and how. A common thread throughout many implementations is temporal dynamics: in other words, what happens to a particular variable over simulated 'time'. This is usually quantified in interactions, but may also be described in terms of broader 'timesteps', which in population-based simulations are often defined as $t = M$ interactions, where M is the population size (Baronchelli et al. 2008), with the general rationale that this resolution gives each agent an opportunity to act as speaker and hearer within a timestep.

How time is conceptualized within a model is important for making clear decisions about the temporal resolution of measurement, which will be determined in large part by the population size. Because each agent is involved in every interaction where there is only one dyad, measuring the mappings of each agent at every timestep may give an interesting picture of the temporal dynamics of learning and coordination within the dyad. However, for models which consider larger populations, measuring every learner at every interaction is ineffective: only two agents within the population will be involved in any given interaction, and not much overall change is likely—thus, a resolution of $t \geq M$ interactions for measurement is more efficient.

As important as *when* you measure is *what* you choose to measure, and often a given model might entail measuring multiple things. These can generally be divided into measures involving communicative success, measures involving agent production, and measures involving agent mappings or internal states. For example, within a single interaction of the simple LSG, we can measure whether a given interaction involved a

9780192886491_Book.indb 311 27-Mar-25 23:03:32

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

successful signal–meaning (i.e. action–state) mapping, what signal *S* produced for the chosen state, and/or the internal action–state mappings of each agent (and the extent to which these overlap). While these measures may seem somewhat redundant, their convergence and temporal dynamics are important. For example, a single interaction can be successful as a matter of a randomly produced signal even where internal mappings are not yet aligned; likewise, brief periods of convergence are less compelling than more temporally stable communicative success.

These basic measurements, in turn, can be used to quantify signalling systems in more complex ways. For example, the entropy of mappings over a temporal window or a population can signal diachronic and/or population stability: in other words, when entropy is low, mappings are predictable and have stabilized within a dyad or population over time, or across a population in a snapshot (Cuskley et al. 2018). Other information theoretic measures (e.g. conditional entropy; Ohmer, Duda, et al. 2022) and measures of structure often applied to signal–meaning mappings can also be used (e.g. the Mantel Test; Mantel 1967, although see Guillot and Rousset 2013), especially where these may make for useful comparison with measures of experimental data (Cuskley 2019; Cuskley et al. 2019; Ferdinand et al. 2019).

Overall, a key concept for these models is the notion of multiple runs (for implementations with 'simple' agents minimally hundreds, but potentially a lot more), which can eliminate the need for inferential statistics common in experimental approaches. In dyadic communication games with human participants, practical limitations mean it is only feasible to gather data from perhaps tens of dyads, who interact for a specified period of time that may be relatively brief. Data produced in these interactions is then compared across dyads and often subjected to inferential statistics to draw more general conclusions. Simulations with simple agents, in contrast, allow us to 'run' a large number of dyads with identical starting parameters, exhaustively sampling the space of possible outcomes. In other words, instead of inferring quantities of interest (e.g. the number of interactions to convergence) from a small sample, we can effectively *generate* their whole distribution. More computationally complex implementations involving, for example, deep neural network agents (for an overview, see Lazaridou and Baroni 2020), and/or large or expanding populations (Cuskley et al. 2018) can make exhaustive sampling of the possibility space less practical. Still, computing detailed descriptive statistics of relevant quantities across multiple runs is generally more effective than subjecting results to (particularly frequentist) inferential statistics.

17.3.2 General strengths and weaknesses

Like any methodological approach to the kinds of broad questions at the heart of language evolution, agent-based models which revolve around communicative interaction have specific strengths and weaknesses. In general, the use of these methods should be motivated by specific questions rooted in investigations of real-world phenomena.

9780192886491_Book.indb 312 27-Mar-25 23:03:32

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Analytical solutions can only be determined for the simplest cases of the implementations described in section 17.2 (see Baronchelli 2016 for the NG and Huttegger 2007 for the LSG). For most interesting cases, simulations are necessary, and increasingly complex questions—such as how perceptual biases and representations interact with emergent communication—lead to increasingly complex simulations. Still, researchers looking to apply this method should be wary of building in *too much* complexity. In other words, a great strength of this approach is its ability to strip away complexity to begin to understand how specific elements of linguistic interaction contribute to the emergence of effective shared communication systems. It can be tempting to treat these methods as a potential way to observe the *entirety* of language evolution—or even some complex subset of it—under controlled conditions, as a kind of closed terrarium for language evolution on a small scale. However, we should not expect agent-based models to be able to recreate large, multi-layered language systems; they are Petri dishes, not a terrarium.

In short, agent-based models can be used to explore how particular factors ranging from learner biases to population structure or world structure affect modelled language systems and language learners. However, the method is not suitable for looking at all of these at once—it cannot provide a detailed simulation of the *emergence* of complex human-like language among interacting agents.[4] Even if the computing power required to do something like this were trivially available, the emergent effects would be too difficult to understand with too many factors interacting at once (Woensdregt et al. 2021). Moreover, expanding the number of varying parameters will explode the possible solution space, making it unlikely that it can be effectively sampled. As is arguably the case with many modern scientific approaches (van Rooij and Baggio 2021), agent-based modelling in language evolution should be applied with principled and explicit mechanistic theories about how factors interact.

Overall, this is as much a limitation of the approach as a feature: the necessity for simplifying assumptions makes these kinds of models ideal complements to observational data and experiments, which are inherently noisier and more complex. These kinds of agent-based models can provide stripped-back information about learner and population dynamics that are impossible to observe effectively or ethically in the 'wild' or in the lab. This makes it a good fit for examining dynamics in a way that abstracts away from phenomena in specific languages. For example, models can be used to probe the extent to which historical change in a specific language (e.g. changes to past tense inflection in English; Cuskley et al. 2014) reflects an idiosyncratic history, or dynamics

[4] Note this is distinct from the *acquisition* of a complex language system given input from human language: there are agent-based models which use fluid construction grammar (FCG) to show that agents can develop rich linguistic representations given input from e.g. English, and a rich sensory environment (Beuls and van Eecke 2023). While not *de novo* emergence, acquisition in FCG agents is significantly more informative for language evolution than, for example, LLMs such as GPT, which learn exclusively from text and show some linguistic *competence*, but lack the kinds of transparent representations acquired by FCG agents.

9780192886491_Book.indb 313 27-Mar-25 23:03:33

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

which may play a role in language change more broadly (Cuskley et al. 2018; see also Pijpops et al. 2015).

17.4 Extensions and other applications

The previous sections focused on simple implementations of the NG and LSG, with a more complex case study of their use in exploring the relationship between perception and language. However, the affordances of dyadic agent-based communication games extend well beyond this. Below, we provide a brief overview of some other applications of these methods for investigating language contact and pragmatics to illustrate their versatility.

Contact linguistics studies the effects of contact between two or more populations that speak different languages. Here, inventories often represent the presence or absence of an entire language, rather than specific signal–meaning mappings. For example, Jansson et al. (2015) and Tria et al. (2015) extend the basic NG to study the dynamic processes underlying the emergence of creole languages, integrating real demographic census data to understand historical creole emergence. The LSG has also been extended to simulations with multiple agents to study contact linguistic phenomena (e.g. Harding Graesser et al. 2019). In line with the generally smaller scale of the LSG, these simulations use more complex agents to communicate about more complex meanings, but the total number of agents is small (only thirty agents compared to > 1,000 [Jansson et al. 2015] and 10,000 [Tria et al. 2015] in the NG approaches). Agent-based approaches have also been used to model language emergence in conditions of contact between Deaf and hearing communities, providing insights into the conditions under which robust community sign languages emerge and persist (Mudd et al. 2020).

Agent-based modelling can also be extended to examine issues in pragmatics, in other words, ways in which meaning plays a role in linguistic interaction, but is not necessarily contained entirely within a linguistic signal. Referencing objects at different levels of specificity (e.g. 'Fido', 'dalmatian', 'dog', 'animal') is a pragmatic phenomenon, and the context—next to other factors such as object typicality or cost of utterance alternatives—plays an important role in determining specificity (e.g. Graf et al. 2016; Hawkins et al. 2018). Recent deep learning approaches to the LSG integrate pragmatic context, working with a contextualized reference game where the target is presented next to a set of distractors. Among others, such setups have been used to explicitly study the formation of specific vs. generic references based on the context (Mu and Goodman 2021; Ohmer, Duda, et al. 2022). Unlike the basic LSG, the inclusion of context can lead to stable systems where signals alone are ambiguous, but communication is nonetheless successful because contextual information resolves ambiguity in any given interaction (Mühlenbernd et al. 2022). Ambiguity can also be resolved by imbuing agents with more sophisticated pragmatic perspective taking (Blokpoel et al. 2019).

9780192886491_Book.indb 314 27-Mar-25 23:03:33

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

17.4.1 Conclusion

Taken together, the examples outlined here demonstrate how changes to agents' 'cognition' (memory constraints, reasoning about the context), population structure (multiple populations of different sizes, specific social network topologies), or the meanings (languages, perceptions, concepts) can be used to adapt communication games to specific research questions. Questions concerning language contact, pragmatics, language, and perception, and rule systems in language are just a small sample of the kinds of phenomena that agent-based models can be used to address (see also e.g. Gong et al. 2014; Lipowska and Lipowski 2022). Here, we have provided a basic introduction to implementing agent-based simulations to explore how effective communication emerges between interacting agents. Many areas remain unexplored from this perspective, giving these models significant potential for expanding our understanding of a wide range of issues in language evolution. In particular, these kinds of models can be compared (or enriched) with corpus or observational data and/or more controlled experimental data, triangulating evidence to better understand the evolution of language.

References

Baronchelli, A. (2016). 'A gentle introduction to the minimal naming game', *Belgian Journal of Linguistics* 30(1): 171–192.

Baronchelli, A., Gong, T., Puglisi, A., and Loreto, V. (2010). 'Modeling the emergence of universality in color naming patterns', *Proceedings of the National Academy of Sciences* 107(6): 2403–2407.

Baronchelli, A., Loreto, V., and Steels, L. (2008). 'In-depth analysis of the naming game dynamics: The homogeneous mixing case', *International Journal of Modern Physics C* 19(05): 785–812.

Barrett, J. A., Cochran, C. T., Huttegger, S. M., and Fujiwara, N. (2017). 'Hybrid learning in signalling games' *Journal of Experimental & Theoretical Artificial Intelligence* 29(5): 1119–1127.

Barrett, J. A. and Zollman, K. J. S. (2009). 'The role of forgetting in the evolution and learning of language', *Journal of Experimental & Theoretical Artificial Intelligence* 21(4): 293–309.

Berlin, B. and Kay, P. (1969). *Basic Color Terms: Their Universality and Evolution*. Berkeley, CA: University of California Press.

Beuls, K. and van Eecke, P. (2023). 'Fluid construction grammar: State of the art and future outlook', in C. Bonial and H. Tayyar Madabushi (eds.), *Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs+NLP, GURT/SyntaxFest 2023)*, 41–50. Washington, DC: Association for Computational Linguistics, https://aclanthology.org/2023.cxgsnlp-1.6/.

Blokpoel, M., Dingemanse, M., Woensdregt, M., et al. (2019). 'Pragmatic communicators can overcome asymmetry by exploiting ambiguity', *Open Science Framework preprint*, doi: 10.31219/osf.io/q56xs.

Castellano, C., Fortunato, S., and Loreto, V. (2009). 'Statistical physics of social dynamics', *Reviews of Modern Physics* 81(2): 591.

9780192886491_Book.indb 315 27-Mar-25 23:03:33

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Catteeuw, D., De Beule, J., Manderick, B., et al. (2011). 'Roth-Erev learning in signaling and language games', *Proceedings of the Benelux Conference on Artificial Intelligence* 23: 67–74.

Centola, D. and Baronchelli, A. (2015). 'The spontaneous emergence of conventions: An experimental study of cultural evolution', *Proceedings of the National Academy of Sciences* 112(7): 1989–1994.

Cuskley, C. (2019). 'Alien forms for alien language: Investigating novel form spaces in cultural evolution', *Palgrave Communications* 5(1): 1–15.

Cuskley, C. and Kirby, S. (2013). 'Synesthesia, cross-modality, and language evolution', *Oxford Handbook of Synaesthesia* 20: 869–907.

Cuskley, C., Pugliese, M., Castellano, C., et al. (2014). 'Internal and external dynamics in language: Evidence from verb regularity in a historical corpus of English', *PLOS ONE* 9(8): e102882, doi: 10.1371/journal.pone.0102882.

Cuskley, C., Loreto, V., and Kirby, S. (2018). 'A social approach to rule dynamics using an agent-based model', *Topics in Cognitive Science* 10(4): 745–758.

Cuskley, C., Dingemanse, M., Kirby, S., and Van Leeuwen, T. M. (2019). 'Cross-modal associations and synesthesia: Categorical perception and structure in vowel–color mappings in a large online sample', *Behavior Research Methods* 51: 1651–1675.

Dingemanse, M., Roberts, S. G., Baranova, J., et al. (2015). 'Universal principles in the repair of communication problems', *PLOS ONE* 10(9): e0136100.

Ferdinand, V., Kirby, S., and Smith, K. (2019). 'The cognitive roots of regularization in language', *Cognition* 184: 53–68.

Franke, M. (2016). 'The evolution of compositionality in signaling games', *Journal of Logic, Language and Information* 25(3): 355–377.

Goldstone, R. L. and Hendrickson, A. T. (2010). 'Categorical perception', *WIREs Cognitive Science* 1(1): 69–78.

Gong, T., Shuai, L. and Zhang, M. (2014). 'Modelling language evolution: Examples and predictions', *Physics of Life Reviews* 11(2): 280–302.

Graf, C., Degen, J., Hawkins, R. X., and Goodman, N. D. (2016). 'Animal, dog, or dalmatian? Level of abstraction in nominal referring expressions', *Proceedings of the 38th annual meeting of the Cognitive Science Society (CogSci)*, 2261–2266.

Guillot, G. and Rousset, F. (2013). 'Dismantling the mantel tests', *Methods in Ecology and Evolution* 4(4): 336–344.

Harding Graesser, L., Cho, K., and Kiela, D. (2019). 'Emergent linguistic phenomena in multi-agent communication games', *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, 3700–3710. Association for Computational Linguistics.

Hawkins, R. D., Franke, M., Smith, K., and Goodman, N. D. (2018). 'Emerging abstractions: Lexical conventions are shaped by communicative context', *Proceedings of the 40th annual meeting of the Cognitive Science Society (CogSci)*, 463–468.

Huttegger, S. M. (2007). 'Evolution and the explanation of meaning', *Philosophy of Science* 74(1): 1–27.

Huttegger, S. M. and Zollman, K. J. S. (2011). 'Signaling games: Dynamics of evolution and learning', in A. Benz, C. Ebert, G. Jäger, and R. van Rooij (eds.), *Language, Games, and Evolution: Trends in Current Research on Language and Game Theory*, 160–176. Berlin, Heidelberg: Springer.

Huttegger, S. M., Skyrms, B., Tarrès, P., and Wagner, E. (2014). 'Some dynamics of signaling games', *Proceedings of the National Academy of Sciences* 111(supplement 3): 10873–10880.

9780192886491_Book.indb 316 27-Mar-25 23:03:33

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Jansson, F., Parkvall, M., and Strimling, P. (2015). 'Modeling the evolution of creoles', *Language Dynamics and Change* 5(1): 1–51.

Lazaridou, A. and Baroni, M. (2020). 'Emergent multi-agent communication in the deep learning era', *arXiv preprint, arXiv:abs/2006.02419*.

Lewis, D. (1969). *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press.

Lipowska, D. and Lipowski, A. (2022). 'Emergence and evolution of language in multi-agent systems', *Lingua* 272: 103331.

Lipowski, A. and Lipowska, D. (2009). 'Language structure in the n-object naming game', *Physical Review E* 80(5): 056107.

Lu, Q., Korniss, G., and Szymanski, B. K. (2009). 'The naming game in social networks: Community formation and consensus engineering', *Journal of Economic Interaction and Coordination* 4(2): 221–235.

Mantel, N. (1967). 'The detection of disease clustering and a generalized regression approach', *Cancer Research* 27(2): 209–220.

Markman, A. B. and Makin, V. S. (1998). 'Referential communication and category acquisition', *Journal of Experimental Psychology: General* 127(4): 331–354.

Mu, J. and Goodman, N. (2021). 'Emergent communication of generalizations', in M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (eds.), *Advances in Neural Information Processing Systems* (Vol. 34), 17994–18007. Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2021/file/9597353e41e6957b5e7aa79214fcb256-Paper.pdf

Mudd, K., De Vos, C., and De Boer, B. (2020). 'An agent-based model of sign language persistence informed by real-world data', *Language Dynamics and Change* 10(2): 158–187.

Mühlenbernd, R., Wacewicz, S., and Z˙ywiczyński, P. (2022). 'The evolution of ambiguity in sender-receiver signaling games', *Games* 13(2): 20.

Ohmer, X., König, P., and Franke, M. (2020). 'Reinforcement of semantic representations in pragmatic agents leads to the emergence of a mutual exclusivity bias', *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society (CogSci)*, 1779–1785.

Ohmer, X., Duda, M., and Bruni, E. (2022). 'Emergence of hierarchical reference systems in multi-agent communication', in N. Calzolari et al. (eds.), *Proceedings of the 29th International Conference on Computational Linguistics (COLING)*, 5689–5706.

Ohmer, X., Marino, M., Franke, M., and König, P. (2022). 'Mutual influence between language and perception in multi-agent communication games', *PLOS Computational Biology* 18(10): 1–28.

Oliphant, M. and Batali, J. (1997). 'Learning and the emergence of coordinated communication', *Center for Research on Language Newsletter* 11(1): 1–46.

Pijpops, D., Beuls, K., and Van de Velde, F. (2015). 'The rise of the verbal weak inflection in germanic an agent-based model', *Computational Linguistics in the Netherlands Journal* 5: 81–102.

Rescorla, M. (2019). 'Convention', in E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (summer edition). Retrieved from https://plato.stanford.edu/ archives/sum2019/entries/convention/.

Rosch, E., Mervis, C. B., Gray, W. D., et al. (1976). 'Basic objects in natural categories', *Cognitive Psychology* 8(3): 382–439.

Schelling, T. C. (1960). *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Scott-Phillips, T. C., Kirby, S., and Ritchie, G. R. (2009). 'Signalling signalhood and the emergence of communication', *Cognition* 113(2): 226–233.

9780192886491_Book.indb 317 27-Mar-25 23:03:33

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Shannon, C. E. (1948). 'A mathematical theory of communication', *The Bell System Technical Journal* 27(3): 379–423.

Skyrms, B. (2010). *Signals: Evolution, Learning, and Information*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199580828.001.0001

Spike, M. (2018). *Minimal Requirements for the Cultural Evolution of Language* (Unpublished doctoral dissertation). Edinburgh: University of Edinburgh.

Steels, L. (1995). 'A self-organizing spatial vocabulary', *Artificial Life* 2(3): 319–332.

Steels, L. and Belpaeme, T. (2005). 'Coordinating perceptually grounded categories through language: A case study for colour', *Behavioral and Brain Sciences* 28(4): 469–488, doi: 10.1017/S0140525X05000087.

Steels, L. and Loetzsch, M. (2012). 'The grounded naming game', *Experiments in Cultural Language Evolution* 3: 41–59.

Suffill, E., Branigan, H., and Pickering, M. (2019). 'Novel labels increase category coherence, but only when people have the goal to coordinate', *Cognitive Science* 43(11): e12796.

Tria, F., Servedio, V. D., Mufwene, S. S., and Loreto, V. (2015). 'Modeling the emergence of contact languages', *PLOS ONE* 10(4): e0120771.

van Eecke, P., Beuls, K., Botoko Ekila, J., and Rădulescu, R. (2022). 'Language games meet multi-agent reinforcement learning: A case study for the naming game', *Journal of Language Evolution* 7(2): 213–223.

van Rooij, I. and Baggio, G. (2021). 'Theory before the test: How to build high-verisimilitude explanatory theories in psychological science', *Perspectives on Psychological Science* 16(4): 682–697.

Verma, G., Swami, A., and Chan, K. (2012). 'The effect of zealotry in the naming game model of opinion dynamics', *MILCOM 2012–2012 IEEE Military Communications Conference*, 1–6.

Woensdregt, M. S., Spike, M., de Haan, R., et al. (2021). 'Why is scaling up models of language evolution hard?', *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 43).

Wolff, P. and Holmes, K. J. (2011). 'Linguistic relativity', *WIREs Cognitive Science* 2(3): 253–265.

9780192886491_Book.indb 318 27-Mar-25 23:03:33

CHAPTER 18

COMPUTATIONAL METHODS FOR LANGUAGE VARIATION AND CONVERGENCE

KATIE MUDD AND BART DE BOER

18.1 Introduction

Evolving systems need variation, as without it selection would have nothing to work with. Language is one example of an evolving system (Steels 2000), which exhibits variation at all linguistic levels. The variant an individual in the population selects is not random, but is conditioned by a host of factors (e.g. sociolinguistic, semantic, and phonological; e.g. Givón 1985). At the lexical level, for example, English speakers may use the word *elevator* or *lift*, and their choice of lexical variant is typically predicted by their location. One method to study how and why variation spreads in a population is using computational modelling. The field of language evolution has a rich history of using computational models; many of these investigations have focused on the processes which lead to convergence. In this chapter, we focus rather on models studying what processes affect the degree and patterning of variation in a population.

The chapter is structured as follows: we first situate variation and convergence research within language evolution research. Following this, the majority of this chapter details a recent model in the field of language evolution which operationalizes a theory about the interaction between cultural context and lexical variation in language emergence (Mudd et al. 2022a), with the aim of providing an example of how a model of linguistic variation can be designed and implemented. After following the in-depth example, we zoom out and list broadly the components of these types of models. To conclude the chapter, we catalogue existing models about variation and convergence in the field of language evolution.

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

18.1.1 Contextualizing variation and convergence

Variation and *convergence* can be placed at two ends of the same spectrum because they both represent the proportion of the population that is using a given linguistic feature. We can also discuss how the variation, if any, is distributed among the population. Though sometimes used interchangeably, it should be noted that convergence is different from *conventionalization*. Rather, conventionalization refers to the use of conventions or a feature that is accepted in a population; a linguistic feature may be conventionalized without the population having converged on the use of that feature (for more on this, see Mudd 2022: 248–249).

In the field of language evolution, computational models have investigated how populations characterized by a high degree of unpredictable variation converge over time (e.g. Steels 1995, using the naming game paradigm detailed in depth in Ohmer and Cuskley, Chapter 17, this volume), and how different population configurations may accelerate or slow down this process. As such, the focus has largely been on how a population with an initially random, or unstructured linguistic system moves towards convergence. Related to this, an early agent-based model by Kirby (2000) showed how language learners move away from holistic form–meaning mappings towards the use of compositional structure. Other methods, like experiments and naturalistic data are sometimes used in tandem with computational methods. There has also been a focus on the regularization of unpredictable variation. Experimental results show that children tend to regularize more than adults, and therefore the absence of unpredictable variation in language may be strengthened by biases in child language learning (Hudson Kam and Newport 2005, 2009). However, while many of these studies do look at the role of variation, they study how variation in the input affects linguistic structure; their goal is seldom to study variation between individuals in the population.

Despite little contact between the fields of language evolution and sociolinguistics, there is undeniable overlap in interests and research questions, albeit at times framed differently (Roberts and Sneller 2020). The field of sociolinguistics predominantly focuses on language variation and change. Following Labov's (1963) seminal study, there has been research into how different sociolinguistic factors affect linguistic variation. The focus of these studies is to highlight the variation present and cataloguing how it is distributed across a population.

In more recent years, in the field of language evolution, more attention has been paid to sign languages, as the sign languages existing in the world today are younger than spoken languages (which have had uninterrupted transmission) and hence may provide more of a window into the early linguistic features of a modern language and how they might remain stable or change over time (Meir et al. 2010). It has been hypothesized that the degree of variation present at various linguistic levels is correlated with the social structure of the community (Meir et al. 2012), but this

9780192886491_Book.indb 320 27-Mar-25 23:03:33

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

relationship is not straightforward and requires more research, as shown in a recent comparison of lexical variation in Kata Kolok, Israeli Sign Language, and British Sign Language and community structure (Lutzenberger et al. in press). One research question here is why and how communities retain the variation present in language emergence. A host of recent computational models, often informed by studies from sign language linguistics, have been used to investigate the relationship between social structure and linguistic variation in sign languages (e.g. Richie et al. 2014; Thompson et al. 2020; Mudd et al. 2022a).

In the following section, we describe in depth one such model, pulling from variation research in the aforementioned fields. We aim to use this example to introduce key concepts about agent-based modelling and illustrate the process of designing one.

18.2 Example model from Mudd et al. (2022a)

In this section, we describe a model by Mudd et al. (2022a) which studies lexical variation in (sign) language emergence. We focus on design decisions and list major discussion points that arose while creating the model, with the aim of providing a 'behind the scenes' view into our agent-based modelling approach. Generally, it is impossible to design model components independently, as models operationalize the relationship between aspects of language. In designing a model, there is back-and-forth between designing one part and another to determine how abstracted each component can be. This model was implemented in Python using the package Mesa (other common implementation methods are discussed in section 18.3.5). The code for this agent-model is available here: https://doi.org/10.6084/m9.figshare.15163872.v1.

18.2.1 The research question

Our goal was to operationalize a hypothesis from Tkachman and Hudson Kam (2020: 226) detailing the relationship between shared context and lexical variation. They put forward that linguistic convergence is only required without shared context, as shared context allows individuals to make use of iconic form–meaning mappings that are shared throughout the population. As such, our goal in creating the model was to study how specific processes affect lexical variation, though models can be created for a host of reasons (see Epstein 2008). More specifically, we would like to model how the different parts of their theory interact and observe the effect on lexical variation. This allows us to see if the proposed relationship between shared context and lexical variation is theoretically sound.

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

18.2.2 Model components

First, we identified the parts of Tkachman and Hudson Kam's (2020) theory that are required to model the relationship between lexical variation and shared context. The crucial components we identified are lexical variation, shared context, and population structure.

In this section, we 1) define each component, 2) outline the aspects of the component that need to be modelled, 3) explain for each component how they interact with other components and how this affects designing each component, and 4) mention discussions we had about each component when designing the model. In representing each of these components, the goal is to sufficiently represent the crucial aspects of each and allow for their interaction, while keeping this representation as simple as possible (see Smaldino 2017 for a discussion of simplicity in models).

18.2.2.1 Shared context

Shared context refers to how much social and psychological knowledge is shared between two or more individuals. As such, modelling shared context requires the representation of some feature(s) specific to an individual's experience for a given concept. Extending this across individuals, what is therefore required for the model is a concept that agents communicate about, plus some representation of how much knowledge they share about that concept.

Major discussions we had when designing the shared context model component:

- How to represent an individual's experience with a concept?
- Should there be overlap between concepts and across individuals (in different groups)?

18.2.2.2 Lexical variation

Lexical variation refers to how word-level preferences vary across a group of individuals. Thus, modelling lexical variation requires a representation of individual preferences for lexical items. Moreover, the interaction between lexical variants and shared context in Tkachman and Hudson Kam's (2020) theory depends on iconicity: as iconicity is subjective and requires shared context, when there is enough shared context between interlocutors, individuals can rely on iconicity to successfully communicate. In other words, without shared context, iconic mappings will not be successfully communicated between individuals. The model, therefore, also needs to represent shared context and iconicity.

How to implement a language game which represents these components in the simplest manner? Minimally, the agent's linguistic representation requires three components: 1) forms, 2) concepts, and 3) what we call culturally salient features, referring to an agent's subjective experience with a concept. In typical language game models (see Ohmer and Cuskley, Chapter 17, this volume), agents play an interaction game and keep track of forms (or signals) and concepts (or meanings). In order for our interpretation of Tkachman and Hudson Kam's (2020) theory to be modelled, we additionally need a link representing an iconic mapping between forms and concepts and we therefore decided to add a representation of the agent's experience with a concept as a separate component in the language

9780192886491_Book.indb 322 27-Mar-25 23:03:33

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

representation: the *culturally salient features* associated with a concept. This way, the iconic mapping is represented by the similarity between the culturally salient features and the form associated with each concept.

Once we determined that language in the model could be represented by these three components, we designed the language game. Tkachman and Hudson Kam's (2020) theory puts forward that 'lexical convergence is required only when communication is impeded without it' (2020: 226). Therefore, in the language game, agents must have the possibility to successfully communicate about a concept without using the same form. This is what led us to design the language game in such a way that individuals can 1) memorize the form of words corresponding to concepts, 2) rely on the iconic mapping between forms and the features of concepts (making use of shared context), or 3) if communication fails, update their forms in order to successfully communicate, as illustrated in Figure 18.1. Note that steps 1 and 3 are similar to the steps of the classic language game (e.g. Steels 1995; Baronchelli 2016).

Given that this model concerns a theory about how lexical variation evolves in the early stages of sign languages, we paid special attention to how the linguistic representation is initialized. The forms of the agents were initialized by the strength of their mapping with the culturally salient features, determined by the parameter *initial_degree_of_overlap*.

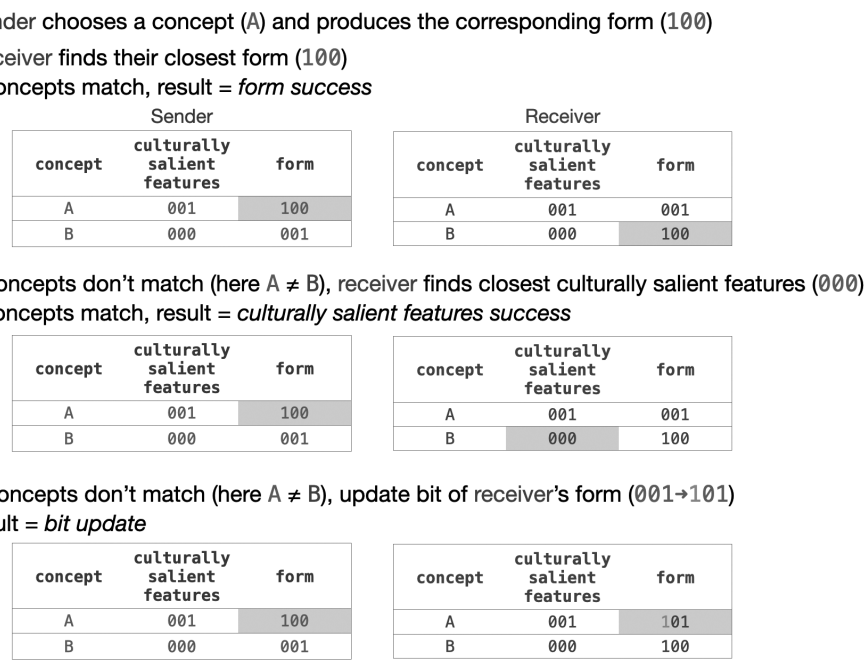


FIGURE 18.1 The steps of the language game with an example.

The sender selects a concept and sends the corresponding form to the receiver whose task it is to return the concept related to the sender's form. To do this, the receiver finds their closest form. If the receiver's concept for their chosen form does not match the sender's concept, the receiver finds their closest culturally salient features to the sender's form (modelling the use of iconicity). If the receiver's concept for their chosen culturally salient features does not match the sender's concept, the receiver updates their form corresponding to the concept the sender initially selected.

Source: Mudd et al. (2022a).

9780192886491_Book.indb 323 27-Mar-25 23:03:33

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Major discussions we had when designing the lexical variation model component:

- How to represent lexical variants in this model?
- How to initialize the forms of agents in the model?
- How and when should forms get updated? Should they be updated if communication is successful or unsuccessful? And should the entire form be updated or just part of the bit string? Should the sender, receiver, or both update their form?
- In the current version of the model, agents can only store a single variant per concept. How many variants should agents be able to store? What we see in many sign languages is that individuals productively use different variants or that perceptually they understand different variants. If agents aren't able to store multiple variants, can we accurately formalize the relationship between shared context and lexical variation?

18.2.2.3 Population structure

Population structure refers to how a population is subdivided in terms of their features. Here, the degree of shared context (and its effect on lexical variation) is manipulated at the population level. To operationalize this, agents were assigned to different groups determining the culturally salient features that correspond to concepts. This parameter (*n_groups*) varies the likelihood that all agents are assigned to one group (*n_groups* = 1) to all agents being assigned to different groups (*n_groups* = number of agents in model). In the version of the model we describe here, at each time step, every agent randomly chooses another agent to play the language game with. In a later version of the model, we used population structure to influence interaction, by studying how lexical variation is affected by agents preferring to interact with agents in their same group (Mudd et al. 2022b).

Major discussions we had when designing the population structure model component:

- In the current model, shared context is determined by the group that participants belong to, such that all agents in one group have the same culturally salient features per concept. However, shared context can vary continuously. For example, agents in the same group could have varying degrees of shared context but more shared between them than with agents from other groups. In this way, there would still be some overlap between the culturally salient features for concepts across groups, but not entirely the same.

18.2.3 Parameter settings and running the model

The main part of this model is the language game, which is repeated over time (for *n_steps*) and played by different pairs of agents. Figure 18.2 shows a schema of the model. As is usual in agent-based modelling, there was first an exploratory stage where ranges for the model parameters were determined, followed by a more detailed investigation of

9780192886491_Book.indb 324 27-Mar-25 23:03:33

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

how these parameters influence the model behaviour. When deciding on which range of values per parameter to run, we typically select values with the shape of an exponential curve in mind. For example, the number of groups (*n_groups*) was run with the values 1, 2, 5, and 10. When a parameter varies between 0 and 1, we try to explore that entire space, by ranging the value from 0 to 1. What we are often interested in is a 'tipping point', the value at which some behaviour changes. For example, in Figure 18.3, it is interesting to note that the curves for when there are five or ten groups is similar, but lower than this (one and two groups), the shape of the curve is quite different. There is no clear guideline or rule on selecting which set of values are selected, therefore transparency in describing models and providing accompanying code is essential.

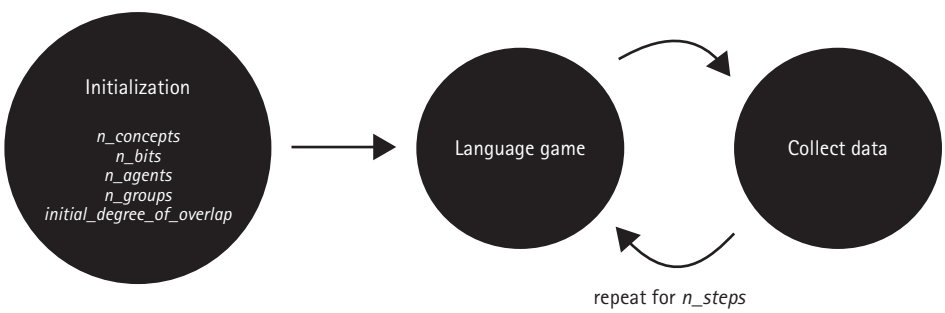


FIGURE 18.2 Visualization of the steps and parameters in the agent-based model.

Source: Mudd et al. (2022a).

Major discussions we had when designing the parameters and model time steps model component:

- What parameters should be fixed vs. explored in a parameter exploration?

18.2.4 Results

To study how lexical variation is affected by different parameters such as the number of groups, which determines the culturally salient features of agents per concept, we investigated if agents produced the exact same forms for a given concept (a lexical distance of 0) or if they produced different forms for a given concept (a lexical distance of 1). Lexical distance is therefore calculated by comparing all pairs of agents across the population. Iconicity is measured at an individual level by comparing each form to its culturally salient features.

The results provide support for Tkachman and Hudson Kam's (2020) theory: in populations with limited context (more groups in the model with different culturally salient features and hence more initial form variation) lexical variability decreases. On the other hand, in populations with much shared context (fewer groups in the model), the initial lexical variability present is nearly fully retained (see Figure 18.3).

9780192886491_Book.indb 325 27-Mar-25 23:03:33

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

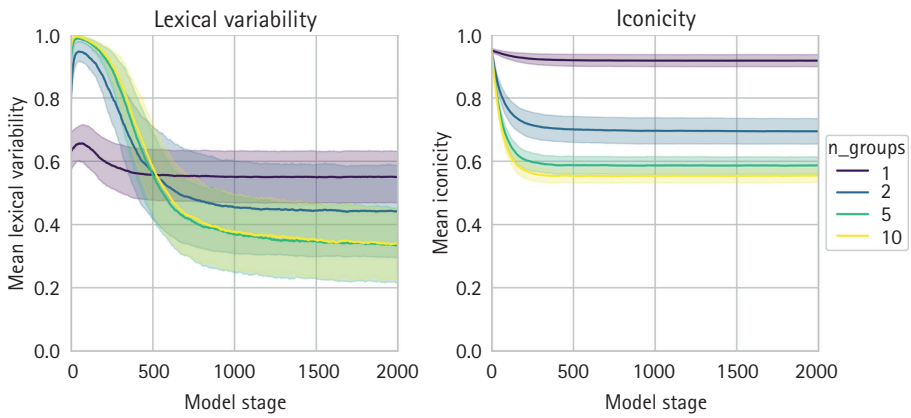


FIGURE 18.3 The mean lexical variability over the 2,000 model stages for a hundred repetitions of a run with ten agents being assigned to different groups depending on the run.

The dark line represents the mean and the shaded area represents the standard deviation of the hundred repetitions. The more groups in the population, the higher the initial lexical variability (at stage 0) but the lower the final lexical variability (at stage 2,000). In addition, when there are fewer groups, the degree of iconicity is higher.

Source: Mudd et al. (2022a).

18.2.5 Discussion: Unexpected dynamics and arbitrary decisions

We would like to conclude this section with a short discussion about certain modelling decisions that were made and how we came to these decisions, to give a more 'behind the scenes' view into model design, implementation, and their effects.

First, we will discuss population size. As shown in Figure 18.3, population size (parameter called *n_agents*) is set at the start of the simulation. Formalizing the relationship between lexical variation and population size was not the focus of this study, but given that population size is a parameter of this model and the subject of related studies on lexical variation (Thompson et al. 2020), we wanted to study its effect. We had not considered how population size, shared context, and lexical variation would interact. Interestingly, when other parameters are held constant, the results of our model were in line with other results about lexical variation and population size: lexical variation is higher in smaller communities. The reason for this is that in larger populations there are simply more initial forms, which instigate a feedback loop consisting of 1) changing forms due to less successful communication, 2) a reduction of iconicity, and 3) fewer language games ending with culturally salient feature success (as shown in Mudd et al. 2022a: fig. 5) This illustrates how having a model allows us to discover unexpected dynamics.

In a similar way, while we had hypothesized that the final amount of lexical variation would be lower in populations with less shared context (more groups in the model), we had not made predictions or considered what the amount of lexical variation would be in language emergence (at the initial stage of the model). As can be seen in Figure 18.4,

9780192886491_Book.indb 326 27-Mar-25 23:03:34

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

with more groups, there is much more initial lexical variability which decreases over time. This is another example of how modelling may draw the attention to phenomena that one might not have considered otherwise.

Finally, we would like to address an initially arbitrary decision made which we decided to keep after the fact. When implementing the update rule for the form (when communication fails), we initially made the decision to flip one bit from the receiver's form to be more similar to that of the sender. After some reflection, we found this to be unrealistic: this would be more akin to a phonetic change than a lexical change, where only one part of the form changes. However, we decided to keep this implementation of the language update because updating one bit of the form is more gradual and probably more representative of the slow changes in language evolution. Changing the entire form would likely cause drastic changes to the system, which we do not observe in real life.

18.3 Modelling linguistic variation

Here, we briefly outline key components in computational models of lexical variation, as shown in the Mudd et al. (2022a) model. Following this, we will list existing models of lexical variation, providing examples of different types of models.

18.3.1 Research questions

Early models focused on the specific processes leading to linguistic convergence (see Ohmer and Cuskley, Chapter 17, this volume) using the Naming Game paradigm (Steels 1995). Another strand of initial computational research in the field of language evolution focused predominantly on systematicity, for example, studying how agents with unstructured, random linguistic systems move towards structured, stable linguistic systems (e.g. Kirby 2000). More recent models have focused on explaining the degree of lexical variation in a community (e.g. Richie et al. 2014; Thompson et al. 2020; Mudd et al. 2022a), with sign language communities in mind. Other models have investigated how network structure affects linguistic variation (e.g. Gong et al. 2011; Josserand et al. 2021).

18.3.2 Typical model components

In section 18.2, we went through the key parts of the Mudd et al.'s (2022a) model, which are more or less representative of what is typically included in a model. We briefly mention these components again here. In a typical agent-based model about variation, agents with a representation of variation exist in a structured population. Agents interact, resulting in changes to their linguistic representation. Agents are endowed with (possibly

9780192886491_Book.indb 327 27-Mar-25 23:03:34

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

biased) learning mechanisms. For instance, agents may generalize or not, and have different memory capacities. In their learning representation, agents may be Bayesian learners (e.g. Griffiths and Kalish 2007; Thompson et al. 2016) or agents may consist of neural models (e.g. Galke et al. 2022). Finally, the signalling space that agents use may also have repercussions on variation; modellers may choose an abstract, discrete, and noiseless signalling space if they are not interested in the effect of the signalling space.

18.3.3 How linguistic variation can be represented in models

Linguistic variation can be represented in many ways. The most minimal way to represent linguistic variation in a population would be to have a binary representation of variation: 0 representing one linguistic feature and 1 representing another linguistic feature. Slightly less minimally, variation could be represented as a bit vector, as in the models by Chater et al. (2009) and Dale and Lupyan (2012), both of which model agents as bit vectors of specified lengths that encode grammatical features (represented by 0s and 1s). These models do not encode meaning, they simply represent variation in form. Other models do include a meaning associated with a form, often using an association matrix of form–meaning mappings (e.g. as used in Smith 2004); a hypothetical example of an association matrix is shown in Figure 18.4. Using these association matrices, agents record how often they observe a form with a meaning, and typically increase the association between successful form–meaning mappings and decrease the association between competing form–meaning mappings. Depending on the research question, the linguistic level can be specified (e.g. at the morphological level in Dale and Lupyan 2012; at the lexical level in Richie et al. 2014) or unspecified (e.g. in the model of Josserand et al. 2021). We suggest using the simplest possible representation necessary to study one's research question.

	🐷	🐴	🐸
/ba/	**80**	9	11
/gi/	12	2	**67**
/du/	8	**89**	22

FIGURE 18.4 Example of an association matrix between forms (rows) and meanings (columns). The numbers correspond to percentages in which a form and meaning are observed together and yield successful communication. Percentages in bold represent the most successful form–meaning mapping. Columns add up to 100.

We recommend that special attention be paid to how forms are initialized in the model: should the forms be randomly generated? Or should the forms be generated

9780192886491_Book.indb 328 27-Mar-25 23:03:34

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

based on another component of the linguistic system? Consider how the initialization of forms maps onto language in the real world. Decisions not only have theoretical implications, but also will affect the model results.

18.3.4 Measuring linguistic variation

Measuring linguistic variation across a population is straightforward: variation can be studied between agents using distance measures. If comparing only one element (e.g. if two agents produce the same lexical item or not), a binary distance measure can be used. When more than one element is being compared, distance measures like Jaccard distance could be used (e.g. as used by Ruland et al. 2020). These distance measures between individuals can then be averaged across the whole population. Additionally, entropy, a measure of variation within a system, is commonly used to measure linguistic variation across a population.

18.3.5 Implementing computational models

Computational models can be implemented in many ways. The earliest agent-based models were made with pen and paper (e.g. Schelling 1971) and any programming language can be used. Several platforms and even dedicated programming languages have been created for agent-based modelling. For example, NetLogo is a programming language with an integrated development environment (IDE) created specifically for agent-based modelling (Wilensky 1999). Second, the Python package Mesa (Kazil et al. 2020) aims to be equivalent to agent-based modelling languages like NetLogo, as it includes built-in tools for visualizing agents and plotting agent-based modelling results, and data collection modules for agents and aspects of the model. Third, the agent-based modelling software Babel is a toolkit for creating and running experiments on emergent communication (Loetzsch et al. 2008; Nevens et al. 2020). It includes a host of modules, including Fluid Construction Grammar and it is written in Common Lisp. For individuals who know how to program, we would recommend implementing the agent-based model in an object-oriented programming language (such as Python or C++), as this is the most flexible option. Regardless of how the model is implemented, it is important that it is well described (see the ODD paradigm for describing agent-based models by Grimm et al. 2010) and that the code is openly available with the research.

18.4 Models of linguistic variation

In this section, we briefly catalogue computational models in the field of language evolution whose focus is explicitly on how one linguistic feature varies, typically in

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

relation to other factors. While numerical simulations have also been used to study linguistic variation (e.g. Baronchelli et al. 2006), we focus here on computational models as they can both represent language with some degree of detail and can allow for interaction between agents, representing the complexity involved in language evolution.

18.4.1 Language games

As discussed in depth by Ohmer and Cuskley (Chapter 17, this volume), the classic language game paradigm, as introduced in the work of Luc Steels (1995), models how conventions emerge in a population of agents that all interact with each other on an equal basis (i.e. all agents can be both speakers and hearers, teachers and learners). These models generally go through a phase of increased variability of the language before converging on a shared set of conventions. However, the focus of these models is on convergence, and the phase of variability is seen as a transitory phenomenon that is not usually investigated further. Recently, this work has been expanded upon using modern deep learning networks (Tieleman et al. 2019).

18.4.2 Network structure

Network structure is an increasingly studied topic in the field of language evolution, with different network configurations being explored in experiments and computational models. For instance, Nettle (1999) has investigated the effect of spatial structure, in combination with social impact theory where in a two-dimensional world, nearby agents have a higher probability of interacting than faraway agents and the strength of influence of one agent on another depends on their social status. He found that differences in social status are essential for new linguistic variants to spread. Josserand et al. (2021) investigate how variation is affected by different network sizes, network topologies, and agent biases in populations where some agents have a particular linguistic bias but consist mostly of agents that do not have that bias (i.e. they can learn any language equally easily). Next, Gong et al. (2011) model the co-evolution of linguistic categorization and lexical items, and model how this is affected by several network types: small world, scale free, fully connected, star, rings, and 2-lattices. A model by Richie et al. (2014) studying lexical convergence and the network structures of signing communities echoes Gong et al.'s (2011) findings: a fully connected network, simulating the Nicaraguan Sign Language community, converges more quickly than a network where not all members are connected, simulating different homesign systems in Nicaragua. These models suggest that the more connected a network is, the more quickly the degree of lexical variation in a population reduces.

9780192886491_Book.indb 330 27-Mar-25 23:03:34

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

18.4.3 Models of sign language communities

Several agent-based models of signing communities have been made to address claims about the factors influencing the degree of lexical variation. As previously mentioned, Richie et al. (2014) compare lexical variation in Nicaraguan Sign Language and homesigners in Nicaragua by modelling different interaction patterns. Further, a model by Thompson et al. (2020) operationalizes the relationship between population size and lexical variation. In their model, agents have a fixed memory size so they can only keep track of a limited number of variants. These agents use hierarchical Bayesian inference such that agents use individual-specific representations (if stored in their memory), but when interacting with an unfamiliar agent, agents will use a general lexical variant. Their model provides support for the theory put forward: smaller populations, allowing for the ability to remember individual-specific lexical variants, allows for lexical variation to be maintained in a population. As discussed in section 18.2, the model by Mudd et al. (2022a) operationalizes the relationship between shared context and lexical variation.

18.4.4 Signal–meaning mapping

The signal space that is used to represent a language influences how easy it is to make iconic form–meaning mappings and this influences how easily conventions and structure emerge. In a computational model, it is easy to change the properties of the form and meaning spaces, for instance, how many dimensions they have (and therefore how easy it is to map a form to the meaning space and vice versa, e.g. de Boer and Verhoef 2012) or what their salient features are.

The model by Mudd et al. (2022a) models how lexical variation is affected by the degree of shared context in a population, which can afford the use of iconic form–meaning mappings. Iconicity in this model, as has been described in depth in section 18.4 is modelled as the overlap between *culturally salient features* (representing the important features of a concept for an individual) and the *form*. Iconicity is an integral part of this model, present both in the initialization of the forms per agent as well as in the language game.

18.5 Conclusion

In this chapter, we have aimed to illustrate that agent-based models are a useful tool to model language variation in a population, as the method models local interactions between language learners which is how language is used. In this chapter, we recommend keeping models as simple as possible, representing only key components of the system

9780192886491_Book.indb 331 27-Mar-25 23:03:34

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

one wants to investigate. When designing models of variation and convergence, we urge readers to consider the effect of modelling decisions, such as deciding how to initialize linguistic forms. By providing an in-depth example, we hope to have shown how the creation of models involves back-and-forth between model components, as well as how models can often lead to unexpected findings.

Acknowledgements

We would like to thank Christine Cuskley for reviewing this chapter and for providing incredibly constructive feedback. In addition, thanks to the Amsterdam Brain and Cognition research unit at the University of Amsterdam, which has supported Katie Mudd with a Talent Grant.

References

Baronchelli, A. (2016). 'A gentle introduction to the minimal naming game', *Belgian Journal of Linguistics* 30(1): 171–192.

Baronchelli, A., Felici, M., Loreto, V., et al. (2006). 'Sharp transition towards shared vocabularies in multi-agent systems', *Journal of Statistical Mechanics: Theory and Experiment* 2006(06): P06014–P06014, https://doi.org/10.1088/1742-5468/2006/06/P06014.

Chater, N., Reali, F., and Christiansen, M. H. (2009). 'Restrictions on biological adaptation in language evolution', *Proceedings of the National Academy of Sciences* 106(4): 1015–1020, https://doi.org/10.1073/pnas.0807191106.

Dale, R. and Lupyan, G. (2012). 'Understanding the origins of morphological diversity: The linguistic niche hypothesis', *Advances in Complex Systems* 15(03n04): 1150017, https://doi.org/10.1142/S0219525911500172.

de Boer, B. and Verhoef, T. (2012). 'Language dynamics in structured form and meaning spaces', *Advances in Complex Systems* 15(03n04): 1150021.

Epstein, J. M. (2008). *Why Model?* [Text.Article]. JASSS, http://jasss.soc.surrey.ac.uk/11/4/12.html.

Galke, L., Ram, Y., and Raviv, L. (2022). 'Emergent communication for understanding human language evolution: What's missing?', *Emergent Communication Workshop at ICLR* 2022, https://openreview.net/forum?id=rqUGZQ-0XZ5.

Givón, T. (1985). *Iconicity in Syntax: Proceedings of a Symposium on Iconicity in Syntax, Stanford, June 24-6, 1983*. Amsterdam: John Benjamins Publishing.

Gong, T., Baronchelli, A., Puglisi, A., and Loreto, V. (2011). 'Exploring the roles of complex networks in linguistic categorization', *Artificial Life* 18(1): 107–121, https://doi.org/10.1162/ARTL_a_00051.

Griffiths, T. L. and Kalish, M. L. (2007). 'Language evolution by iterated learning with Bayesian agents', *Cognitive Science* 31(3): 441–480, https://doi.org/10.1080/15326900701326576.

Grimm, V., Berger, U., Bastiansen, F., et al. (2006). 'A standard protocol for describing individual-based and agent-based models', *Ecological Modelling* 198(1–2): 115–126, https://doi.org/10.1016/j.ecolmodel.2006.04.023.

9780192886491_Book.indb 332 27-Mar-25 23:03:34

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Hudson Kam, C. L. and Newport, E. L. (2005). 'Regularizing unpredictable variation: The roles of adult and child learners in language formation and change', *Language Learning and Development* 1(2): 151–195, https://doi.org/10.1080/15475441.2005.9684215.

Hudson Kam, C. L. and Newport, E. L. (2009). 'Getting it right by getting it wrong: When learners change languages, *Cognitive Psychology* 59(1): 30–66, https://doi.org/10.1016/j.cogpsych.2009.01.001.

Josserand, M., Allassonnière-Tang, M., Pellegrino, F., and Dediu, D. (2021). 'Interindividual variation refuses to go away: A Bayesian computer model of language change in communicative networks', *Frontiers in Psychology* 12, https://www.frontiersin.org/articles/10.3389/fpsyg.2021.626118.

Kazil, J., Massad, D., and Crooks, A. (2020). 'Utilizing Python for agent-based modeling: The Mesa Framework', *Social, Cultural, and Behavioral Modeling* 308–317.

Kirby, S. (2000). *The Evolutionary Emergence of Language: Syntax Without Natural Selection: How Compositionality Emerges from Vocabulary in a Population of Learners*, https://doi.org/10.1017/CBO9780511606441.019.

Labov, W. (1963). 'The social motivation of a sound change', *WORD* 19(3): 273–309, https://doi.org/10.1080/00437956.1963.11659799.

Loetzsch, M., Wellens, P., Beule, J. D., et al. (2008). *The Babel2 Manual.*

Lutzenberger, H., Mudd, K., Stamp, R., and Schembri, A. (in press). The social structure of signing communities and lexical variation: A cross-linguistic comparison of three unrelated sign languages.

Meir, I., Sandler, W., Padden, C., and Aronoff, M. (2010). 'Emerging Sign Languages', in M. Marschark and P. E. Spencer (eds.), *The Oxford Handbook of Deaf Studies, Language, and Education* (Vol. 2), 267–280. Oxford: Oxford University Press.

Meir, I., Israel, A., Sandler, W., et al. (2012). 'The influence of community on language structure: Evidence from two young sign languages', *Linguistic Variation* 12(2): 247–291, https://doi.org/10.1075/lv.12.2.04mei.

Mudd, K. (2022). *How Social Structure Affects the Persistence and Features of Shared Sign Languages*. Brussel: Vrije Universiteit.

Mudd, K., de Vos, C., and de Boer, B. (2022a). 'Shared context facilitates lexical variation in sign language emergence', *Languages* 7(1): 31, https://doi.org/10.3390/languages7010031.

Mudd, K., Van Soom, M., de Vos, C., and De Boer, B. (2022). 'The effect of network structure on lexical variablity in sign language communities', in A. Ravignani, R. Asano, D. Valente, F. Ferretti, S. Hartmann, M. Hayashi, et al. (eds.), *The Evolution of Language: Proceedings of the Joint Conference on Language Evolution (JCoLE)*, 525–527. Nijmegen: Joint Conference on Language Evolution (JCoLE).

Nettle, D. (1999). 'Using social impact theory to simulate language change', *Lingua* 108: 95–117.

Nevens, J., Van Eecke, P., and Beuls, K. (2020). 'A practical guide to studying emergent communication through grounded language games', *ArXiv:2004.09218 [Cs]*, http://arxiv.org/abs/2004.09218.

Richie, R., Yang, C., and Coppola, M. (2014). 'Modeling the emergence of lexicons in homesign systems', *Topics in Cognitive Science* 6(1): 183–195, https://doi.org/10.1111/tops.12076.

Roberts, G. and Sneller, B. (2020). 'Empirical foundations for an integrated study of language evolution', *Language Dynamics and Change* 10(2): 188–229.

Ruland, M., Andirkó, A., Romanowska, I., and Boeckx, C. (2020). 'An agent-based model of the gradual emergence of modern linguistic complexity', *BioRxiv* 380683, https://doi.org/10.1101/2020.11.12.380683.

9780192886491_Book.indb 333 27-Mar-25 23:03:34

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Schelling, T. C. (1971). 'Dynamic models of segregation', *The Journal of Mathematical Sociology* 1(2): 143–186, https://doi.org/10.1080/0022250X.1971.9989794.

Smaldino, P. E. (2017). 'Models are stupid, and we need more of them', in *Computational Social Psychology*, 311–331, https://doi.org/10.4324/9781315173726-14.

Smith, K. (2004). 'The evolution of vocabulary', *Journal of Theoretical Biology* 228(1): 127–142, https://doi.org/10.1016/j.jtbi.2003.12.016.

Steels, L. (1995). 'A self-organizing spatial vocabulary', *Artificial Life* 2(3): 319–332.

Steels, L. (2000). 'Language as a complex adaptive system', in M. Schoenauer, K. Deb, G. Rudolph, et al. (eds.), *Parallel Problem Solving from Nature PPSN VI*, 17–26. Berlin, Heidelberg: Springer, https://doi.org/10.1007/3-540-45356-3_2.

Thompson, B., Kirby, S., and Smith, K. (2016). 'Culture shapes the evolution of cognition', *Proceedings of the National Academy of Sciences* 113(16): 4530–4535, https://doi.org/10.1073/pnas.1523631113.

Thompson, B., Raviv, L., and Kirby, S. (2020). 'Complexity can be maintained in small populations: A model of lexical variability in emerging sign languages', *Proceedings of the 13th International Conference on the Evolution of Language* 3, https://doi.org/doi:10.17617/2.3190925.

Tieleman, O., Lazaridou, A., Mourad, S., et al. (2019). *Shaping representations through communication: Community size effect in artificial learning systems* (arXiv:1912.06208). arXiv, http://arxiv.org/abs/1912.06208.

Tkachman, O. and Hudson Kam, C. L. (2020). 'Measuring lexical and structural conventionalization in young sign languages', *Sign Language & Linguistics* 23(1–2): 208–232, https://doi.org/10.1075/sll.00049.tka.

Wilensky, U. (1999). *Netlogo*. Center for Connected Learning and Computer-Based Modeling, Northwestern University, http://ccl.northwestern.edu/netlogo/

9780192886491_Book.indb 334 27-Mar-25 23:03:34

CHAPTER 19

LANGUAGE EVOLUTION WITH DEEP LEARNING

MATHIEU RITA, PAUL MICHEL, RAHMA CHAABOUNI, OLIVIER PIETQUIN, EMMANUEL DUPOUX, AND FLORIAN STRUB

19.1 Introduction

Social animals have been found to use some means of communication to coordinate in various contexts: foraging for food, avoiding predators, mating, etc. (Hauser 1996). Among animals, however, humans seem to be unique in having developed a communication system, natural language, that transcends these basic needs and can represent an infinite variety of new situations (Hauser et al. 2002) to the extent that language itself becomes the basis for a new form of evolution: cultural evolution. Understanding the emergence of this unique human ability has always been a vexing scientific problem due to the lack of access to the communication systems of intermediate steps of hominid evolution (Harnad et al. 1976; Bickerton 2007). In the absence of data, a tempting idea has been to reproduce experimentally the process of language emergence in either humans or computational models (Steels 1997; Myers-Scotton 2002; Kirby 2002).

Experimental paradigms with humans (Kirby et al. 2008; Raviv et al. 2019; Motamedi et al. 2019) have produced significant insights into language evolution. Still, their scope is limited due to the inability to replicate key aspects of language evolution, such as communication within and across large populations and the study of long evolutionary timescales. Computer modelling can help overcome these limitations and has played a prominent role in studying language evolution for a long time (Lieberman and Crelin 1971). In particular, agent-based modelling has been used from the early days of the language evolution research 'renaissance' (Hurford 1989; Steels 1995) and is still

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

a very active and influential field (Smith et al. 2003; Ke et al. 2008; Reali and Griffiths 2009, 2010; Vogt 2009; Gong et al. 2014; Brace et al. 2015; Kirby et al. 2015; Cuskley et al. 2017).

Meanwhile, in the last decade, the field of machine learning has rapidly developed with the advent of deep learning. Deep neural networks have achieved human-level performance in various domains, including image recognition (He et al. 2016; Chen et al. 2020), natural language processing (Devlin et al. 2018; Brown et al. 2020), automatic translation (Bahdanau et al. 2014; Vaswani et al. 2017), and reinforcement learning (Silver et al. 2016).

This chapter aims to introduce the technical and conceptual background required for using deep learning to simulate language evolution, that is to simulate both the emergence of communication in evolutionary timescales and patterns of language change in historical timescales (Kottur et al. 2017; Lazaridou et al. 2018; Lazaridou and Baroni 2020).

First, we present how to implement a communication game (section 19.2), including formalizing it as a machine learning problem (section 19.2.1), designing neural network agents (section 19.2.2), and making agents learn to solve the game (section 19.2.3). Second, we examine the Visual Discrimination Game (Lewis 1969) as a case study (section 19.3), which has been widely explored in neural emergent communication research. Finally, we provide an overview of recent emergent communication simulations with neural networks, highlighting the successes, limitations, and future challenges (section 19.4).

Supplementary technical materials can be found at https://github.com/MathieuRita/LangageEvolution_with_DeepLearning.

19.2 Designing communication games with deep learning

Communication games (Lewis 1969; Steels 1995; Baronchelli et al. 2010) are a framework used to investigate how perceptual, interactive, or environmental pressures shape the emergence of structured communication protocols (Kirby et al. 2008; Cuskley et al. 2017; Raviv et al. 2019). This framework has primarily been studied over the past fifty years and is still one of the leading simulation frameworks in language evolution. See Chapter 17, this volume for more details. This section presents how to simulate communication games using deep learning. First, we frame the communication game as a multi-agent problem, where each agent is represented by a deep neural network (section 19.2.1). Second, we define communicative agents (section 19.2.2). Third, we use machine learning optimization to train agents to solve the communication game (section 19.2.3).

9780192886491_Book.indb 336 27-Mar-25 23:03:35

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

19.2.1 Framing communication games as a machine learning problem

19.2.1.1 Machine learning is well suited for simulating communication games

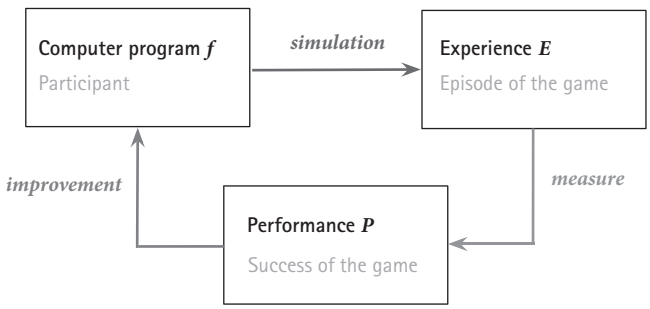


FIGURE 19.1 Iterative learning process in a machine learning problem.

Step 1 (simulation): The computer program *f* performs an experience *E* of the task *T*. Step 2 (measure): The task's success is measured through a performance measure *T*. Step 3 (improvement): Based on its performance, the computer program *f* updates itself to improve its future performance, i.e. learns. Communication games can be framed as a machine learning problem by modelling agents as computer programs. The experience *E* corresponds to an episode of the game, while a performance measure *P* measures the game's success.

Mitchell (1997) defines machine learning as follows:

> 'A computer program f is said to learn from an experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.'

Machine learning is well suited to frame communication games: participants develop a language through trial and error during a communication game. They iteratively adapt their language production and understanding to achieve a given task for which at least one agent lacks information (Tadelis 2013). While game theoretic approaches analyse stable communication protocols (Crawford and Sobel 1982; Skyrms 2010), studying the dynamic learning process is a more challenging and richer problem. Borrowing Mitchell's (1997) notations, this dynamic process can be framed as a machine learning problem where participants are computer programs f that perform the communication game T. The game's success is measured by P after each episode E of the game, and participants update their communication protocol based on the outcome. After enough iterations, the participants may *converge*, i.e. stabilize on a successful communication protocol, allowing them to solve

9780192886491_Book.indb 337 27-Mar-25 23:03:36

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

the game. This iterative learning process is illustrated in Figure 19.1 and is the fundamental idea of machine learning.

19.2.1.2 *Formalizing communication games as a machine learning problem*

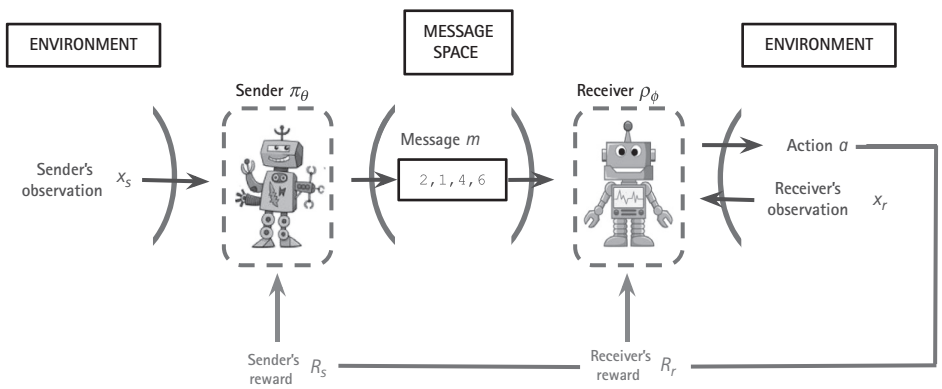


FIGURE 19.2 Scheme of a two-player communication game.

Step 1 (simulation): Agents play a round of the game: (1) both agents get observations from the environment, (2) The sender sends a message to the receiver, (3) The receiver uses the message and its observation to perform an action in the environment. Step 2 (measure): One reward signal per agent measures the game's success. Step 3 (improvement): Agents receive the reward signals and update their behaviour towards better solving the game.

For simplicity, we focus in this chapter on two-player communication games where one agent, the 'sender', sends messages to a second agent, the 'receiver', that parses them and takes actions to solve the task in an environment.[1] Formally, the 'sender' and 'receiver' are parametric models respectively denoted by π_θ and ρ_ϕ with parameters θ and ϕ. Both parametric models will further be designed as deep neural networks. As illustrated in Figure 19.2, a round of the game proceeds as follows:

- The sender π_θ and receiver ρ_ϕ get observations from their environment denoted by x_s and x_r.
- The sender π_θ sends a message m to the receiver ρ_ϕ where m is a sequence of symbols taken from a fixed vocabulary $\mathcal{V}$.
- The receiver ρ_ϕ uses the message m and its observation x_r to perform an action a towards achieving the task.

[1] This setting is referred to as dyadic unidirectional communication games in the literature (Shannon 1948; Harsanyi 1967; Lewis 1969; Cho and Kreps 1987).

9780192886491_Book.indb 338 27-Mar-25 23:03:38

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

The task's success is then measured by two reward signals $\mathcal{R}_s$ and $\mathcal{R}_r$ which are given to the sender π_θ and the receiver ρ_ϕ respectively to improve their protocols. Throughout the game, both agents must agree on a common language to solve the game. Importantly, the emergent language is not defined by explicit language rules but implicitly encoded by the sender's parameters θ.

Remark: This chapter presents a simplified formalism of communication games. Rigorously, communication games should be framed as a special case of Markov Games that provide a broader formal framework for reasoning about multi-agent problems. For further information, refer to (Littman 1994).

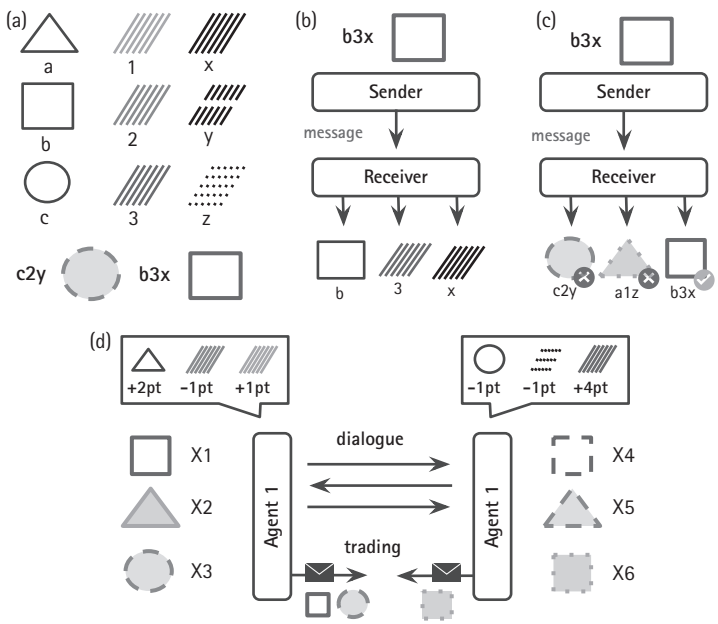


FIGURE 19.3 Examples of Lewis and negotiation games.

(a) Example of Lewis game object's attributes decomposition (shape, colour, style). (b) In the Lewis reconstruction game, the sender observes an object composed of several independent attributes and describes it to the receiver. The receiver must then predict the initial object attributes. (c) In the Lewis discrimination game, the receiver must retrieve the object within a set of distractors. Such a setting does not require manually defining independent attributes, allowing the use of ambiguous real data inputs such as images. (a–c) Such Lewis games usually aim to explore the disentanglement skills of the sender towards producing a compositional language under different scenarios or learning pressures. (d) In negotiation games, agents value objects or attributes differently and get a set of initial objects. They then start dialoguing before executing a final trade. Such tasks involve diverse language interactions such as multiturn communication, non-fully cooperative games, repeated games, or action binding.

9780192886491_Book.indb 339 27-Mar-25 23:03:39

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

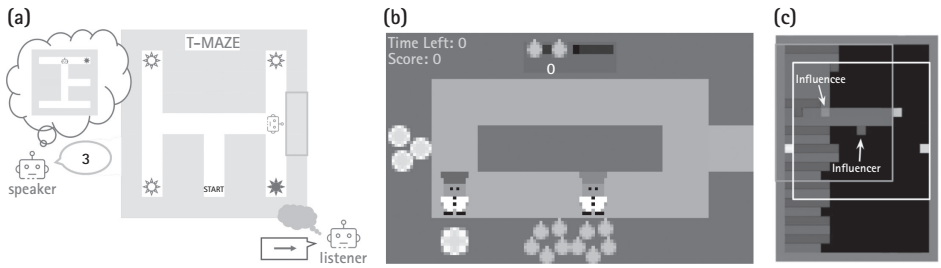


FIGURE 19.4 Attempts to go beyond Lewis and negotiation games by embodying agents into a two-dimensional world.

(a) In the following instruction tasks, the sender is aware of the extensive state of the world and must instruct the receiver on how to reach a pre-defined goal. Importantly, the receiver only has a partial view of its environment. Such tasks aim to explore how basic embodiment properties may shape communication. (b) In coordination games, the agent needs to communicate to execute joint tasks or improve coordination and success through communication. Such tasks ground language to actions. (c) In social dilemma games, agents are surrounded by multiple agents and must behave accordingly to survive. Such tasks explore multichannel communication or behavioural communication through actions.

Sources: (a) Kalinowska et al. (2022); (b) Carroll et al. (2019); (c) Jaques et al. (2019). Reproduced with permission.

In a communication game, the deep neural agents aim to build communication and action policies. This is realized by maximizing their reward. The following is therefore needed:

1. Design the communicative agents as neural networks (section 19.2.2)
2. Train agents to build a shared communication protocol (section 19.2.3)

Figures 19.3 and 19.4 represent communication games commonly studied in language emergence simulations with deep learning. The former presents simple Lewis and negotiation games, while the latter showcases efforts to build more realistic scenarios.

> **Remark:** At the time of writing, many Python libraries, like PyTorch (Paszke et al. 2017) and Jax (Bradbury et al. 2018), are used for easy implementation and optimization of neural networks and are particularly helpful for beginners due to the abundance of online examples.

19.2.2 Designing communicating agents with neural networks

To model communicative agents, we first break them into functional modules that enable interaction with the environment and other agents (section 19.2.2.1). Then, we define neural networks and explain how they can be used to parameterize these functional modules (section 19.2.2.3). Finally, we introduce neural senders and receivers as specific types of neural communicative agents (section 19.2.2.3).

9780192886491_Book.indb 340 27-Mar-25 23:03:39

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

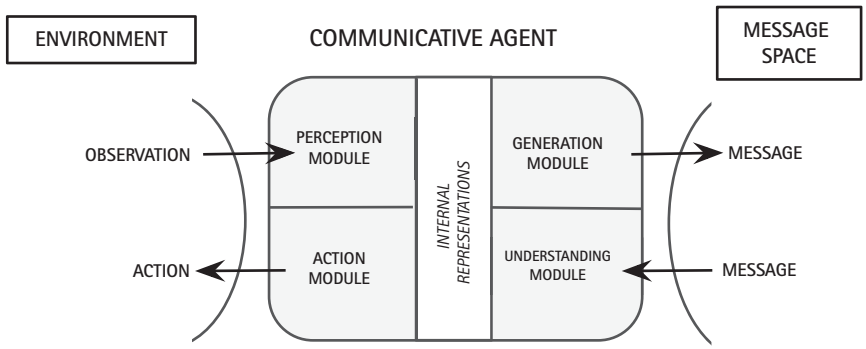


FIGURE 19.5 General view of a communicative agent.

A communicative agent is composed of four functional modules: a perception module that maps an observation to an internal representation; a generation module that maps an internal representation to a message; an understanding module that maps a message to an internal representation; an action module that maps internal representations to an action.

19.2.2.1 Designing a communicative agent as functional modules

As depicted in Figure 19.5, a communicative agent should be able to interact with:

- *Its environment* by either passively observing it or actively taking actions that influence it;
- *Another agent* using a message space by passively receiving or actively sending messages.

Therefore, four functional modules are typically needed to model agents: perception, generation, understanding, and action. (1) The perception module maps an environment's view to an *internal representation*, (2) the generation module generates a message based on internal representations, (3) the understanding module takes a message and builds an *internal message representation*, (4) the action module maps an *internal representation* to an action in the environment.

Neural networks are suited for modelling and combining these modules.

19.2.2.2 Short introduction to neural networks

A neural network f_θ is a parametric model approximating a function or probability distribution based on data. It maps vector inputs to outputs through a succession of linear and non-linear operations. Its learnable parameters θ, called the *weights*, are used to perform the linear operations. The fundamental building block of a neural network is made of two operations:

- A **linear transformation** applying the matrix of weights θ^i to the incoming input;
- A **non-linear transformation** σ, called the *activation function* (typically sigmoid function, hyperbolic tangent (LeCun et al. 1998), or ReLU (Nair and Hinton 2010).

9780192886491_Book.indb 341 27-Mar-25 23:03:40

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

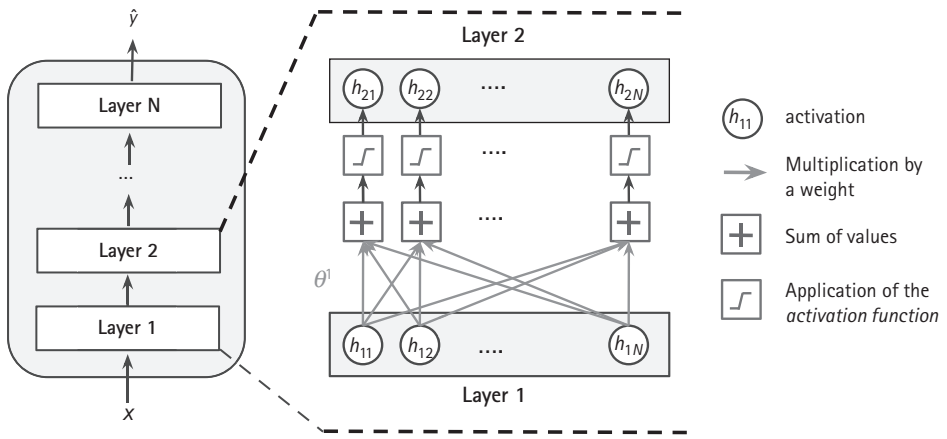


FIGURE 19.6 Scheme of a neural network and the operations between two layers.

A neural network is a function that takes an input $x = (x1, \ldots, xn)$ and maps it to an output $\hat{y}$. It comprises several layers of activations. Each layer results from the application of two operations on the activations of the previous layers: (1) a linear transformation (multiplication by weights and sum), (2) a non-linear transformation (application of a non-linear function). A neural network is parameterized by all the weights acting between each layer.

As displayed in Figure 19.6, these operations are stacked at each *layer*, transforming the input x to a prediction $\hat{y}$ through multiple linear and non-linear transformations.

Remark: Neural networks have a crucial property: all operations are differentiable. This allows for using gradient-based methods to learn the weights (see section 19.2.3).

When training a neural network, the goal is to find the optimal weights θ such that the neural network f_θ accurately maps inputs to their corresponding outputs. Neural networks with enough weights can represent complex functions due to their high expressive power, approximating any continuous function with any level of precision (Hornik et al. 1989). However, computation or data limitations can hinder this process. Deep learning investigates how to adapt networks' architecture or weight matrix shape to overcome these limitations. Figure 19.7 presents the main network architectures and the data they are suitable for.

(a) **Multi-layer perceptron (MLP)** (Rosenblatt 1958) is commonly used to process scalar and heterogeneous input data. It consists of stacked layers, each composed of a linear transformation and a non-linear activation function. A softmax transformation can transform the final activation into a probability distribution.

(b) **Convolutional Neural Network (CNN)** (LeCun et al. 1988) are originally inspired by the visual cortex of animals (Fukushima 1980), and are primarily used for visual data. The CNN's core block consists of three operations: a convolution, a non-linear activation function, and an optional downsampling function,

9780192886491_Book.indb 342 27-Mar-25 23:03:40

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

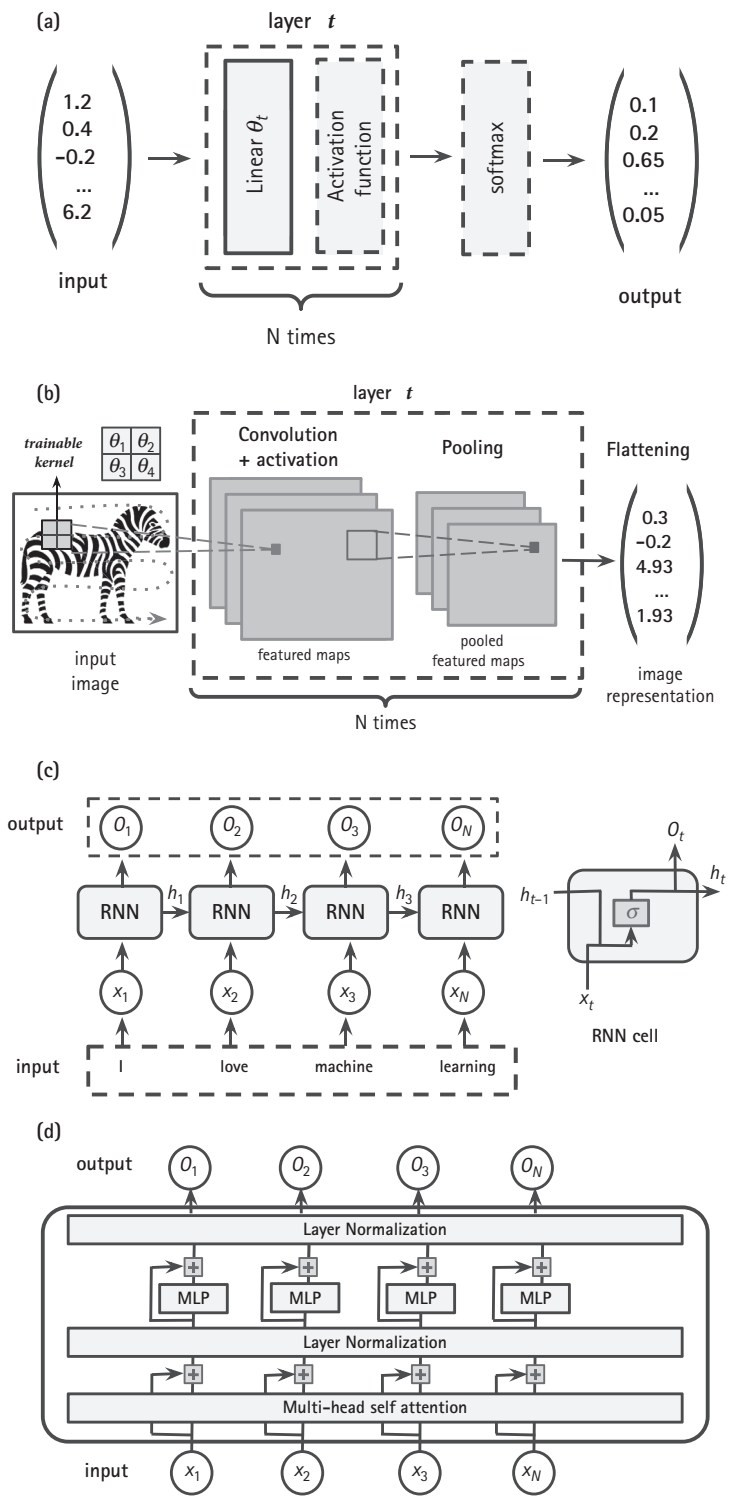


FIGURE 19.7 Classic neural network modules.

Other architectures and/or variants such as Graph Neural Networks (Kipf and Welling 2016) for graphs are not presented here.

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

a.k.a. pooling. The convolution uses a set of filters to scan the input image and detect specific features to build so-called feature maps. The activation function is applied to the feature maps, and then pooling is performed (e.g. taking maximum value over a window of the feature maps). This process is repeated several times, with each subsequent block learning more complex and abstract features. The final layer's output is flattened to provide an abstract input representation. Through the succession of convolutions, the neural network builds a hierarchy of features that capture specific features of the input, e.g. edges and colours.

(c) **Recurrent Neural Network (RNN)** (Elman 1990; Mikolov et al. 2010; Sutskever et al. 2011) is an architecture designed to process input sequences one element at a time while maintaining an internal state that retains information about the past sequence elements. The memory of the RNN is called a state and denoted by h_t. It is updated every time the RNN processes an element, allowing it to use information from the past memory h_{t-1}. The RNN generates an output representation o_t as it processes each new element x_t of the ongoing sentence, and the final output o_N is often used to represent the entire sequence. However, Vanilla RNNs suffer from training instability when working with long sequences. Alternative architectures with more advanced in-cell operations like LSTMs (Hochreiter and Schmidhuber 1997) or GRUs (Chung et al. 2014), and regularization methods like layer normalization (Ba et al. 2016) or gradient clipping (Pascanu et al. 2013) mitigate this issue.

(d) **Transformer** (Vaswani et al. 2017) is a more recent architecture that processes sequences in parallel rather than sequentially using attention mechanisms. These mechanisms allow us to selectively focus on different parts of the input sequence when processing it by differently weighting each input element based on its relevance at a given processing step. See Vaswani et al. (2017) for details on attention. Transformers are generally much faster and more efficient than RNNs. Although less intuitive than RNNs, they are replacing recurrent architectures due to their better performances and scalability advantages. Recent dialogue agent successes are based on the Transformer architecture (Brown et al. 2020; Hoffmann et al. 2022a).

19.2.2.3 Neural functional modules

Several network architectures can be considered when designing agents modules defined in section 19.2.2.1. This section presents some common choices for each module.

Perception module The perception module maps an observation of the environment to an *internal representation*. The choice of architecture depends on the input observation, which differs across games. For example, a *Convolutional Neural Network* (LeCun et al. 1988) is suitable for generating image representations from visual input data, as illustrated in Figure 19.7.

9780192886491_Book.indb 344 27-Mar-25 23:03:42

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Generation module The generation module maps an internal representation, i.e. a vector of a given dimension, into a message. Recurrent neural networks (RNN) (Elman 1990; Mikolov et al. 2010) and Transformers (Vaswani et al. 2017) are well suited for sequences and are hence used in standard emergent communication settings (Kottur et al. 2017; Lazaridou et al. 2018; Chaabouni et al. 2019, 2022; Li and Bowling 2019; Rita et al. 2022a, 2022b). Communication is mainly based on discrete messages, even if some works consider continuous communication protocol (Tieleman et al. 2019).

> **Remark:** To shape the message space, a *vocabulary* of symbols and a *maximum length* must be introduced. It's also possible to add an end-of-sentence token *EoS* to indicate the end of the message. When making these design choices, task complexity should be considered; a larger vocabulary and message length allow for communicating more information/concepts, while a smaller vocabulary and message length require better information compression and, hence, a more structured communication protocol.

Understanding module The understanding module maps a message to an internal representation. Since messages are discrete sequences, RNNs, and Transformers are well suited for this module.

Action module The action module maps an internal representation of an action in the environment. Since the internal representations are scalars and actions a finite set of possibilities, a well-suited architecture is the Multi-Layer Perceptron followed by a softmax that draws a probability distribution over the potential actions.

> **Remark:** Deep learning techniques allow training a system composed of multiple differentiable modules end-to-end. The agent is seen as a single block that provides a prediction given input and output data instead of past methods that glue independently trained/designed blocks together. In communication games, the sender and receiver are both fully differentiable individually. However, the message generation between them does not necessitate *on purpose* to separate the training of the agents. Nonetheless, the message generation can still be made differentiable, as described in section 19.2.3.2.

Good practice: Exploring various neural architectures is a common reflex when starting with deep learning. However, its impact is limited in practice compared to other experimental choices such as task definition, optimization, data, and training objective. Basic neural architectures are recommended to avoid compounding factors when comparing methods.

9780192886491_Book.indb 345 27-Mar-25 23:03:42

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

19.2.2.4 *Modelling neural network communicative agents in communication games*

Section 19.2.2 presents the components of a general communicative agent, though not all modules may be used during a game. Figure 19.8 illustrates sender and receiver modelling in a unidirectional game. This modelling is used in the use case we derive in section 19.3, namely the Visual Discrimination Game.

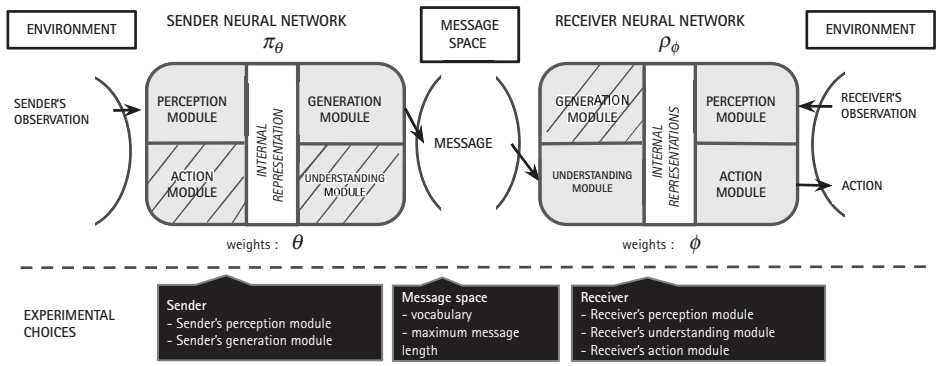


FIGURE 19.8 Summary of neural communicative agents modelling in a unidirectional communication game.

The sender only uses a perception module to process observations and a generation module to create messages. The receiver uses a perception module to process observations, an understanding module to process messages, and an action module to interact with the environment. A vocabulary and maximum message length are defined for the sender's generation. All the experimental choices are framed under the dashed line.

19.2.3 Optimizing the agents to solve the game

In deep learning, the goal is to train neural networks to solve a task, i.e. find the optimal weights that maximize their performance. This section covers optimization techniques for training neural networks and their application to communication games.

19.2.3.1 *Optimizing a machine learning problem*

Data and learning techniques To train neural networks, suitable learning techniques must be chosen depending on the task and the availability of training *data*, which consists of input–output pairs (x, y). Two standard techniques used to solve communication games are:

- **Supervised Learning (SL)**: The neural network is given a training set $\mathcal{D}_{train} = (x_n, y_n)_{n=1}^N$ of N input–output pairs (x_n, y_n), and its objective is to learn how to map inputs to their corresponding outputs. An example of a supervised language task is the translation: the network learns to map one language to another by training on pairs x_n and y_n, where each pair consists of aligned source and target

9780192886491_Book.indb 346 27-Mar-25 23:03:44

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

sentences. Supervised learning finds the weights that enable the network to generalize this mapping to new, unseen examples drawn from the same distribution as the training data, e.g. trying to translate beyond the initial corpus. In communication games, supervised learning tasks often involve classification (e.g. object selection, attribute reconstruction, translation) and regression (e.g. drawing, pixel reconstruction).

- **Reinforcement learning (RL)**: In RL, a neural network, or agent, must perform a sequence of actions to resolve a task within its environment. These actions yield *rewards* that gauge the effectiveness of the network's task performance. The network is then optimized to maximize its expected reward, i.e. performing the sequence of actions that lead to the highest task success. Noteworthy, the probability of action is called a policy in RL. In communication games, the sender produces a sequence of symbols to assist the receiver in completing a predetermined task. If this sequence leads to a successful outcome, the sender is rewarded positively; otherwise, it receives a negative reward. Through iterative trial and error, the sender refines its sequence of symbols towards maximizing its reward and ultimately solving the game, as further detailed in section 19.2.3.2

Supervised learning is easy to apply and highly reproducible but requires a known target. On the other hand, reinforcement learning is more generic and only requires a score to be defined at the cost of being more complex. For instance, to train a network to play chess, supervised learning would involve imitating the moves of a pro-player with a dataset (Silver et al. 2016), while reinforcement learning would require playing the whole game and rewarding victories: the training is more complex and slower, but it does not require data (Silver et al. 2017). It is noteworthy that the reinforcement learning reward can be defined arbitrarily, e.g. one may give an extra bonus when winning the game while preserving the queen, or it could also be used on top of a supervised training regime. This approach has been applied to train large dialogue systems (Ouyang et al. 2022) by imitating the human language and refining it with reinforcement learning.

The loss function: Regardless of the learning technique, the task's success is optimized by introducing a proxy, the *loss function* $l\left(f_\theta; x, y\right)$. The goal is then to find weights θ such that the neural network f_θ minimizes the average loss function $\mathcal{L}_\theta$ over the entire training dataset $\mathcal{D}_{train}$:

$$\min_\theta \mathcal{L}_\theta, \qquad \text{where}: \quad \mathcal{L}_\theta = \mathbb{E}_{(x,y)\sim\mathcal{D}_{train}}\left[l\left(f_\theta; x, y\right)\right] \tag{1}$$

Loss functions vary depending on the network output and the training task (Bishop 2006; Goodfellow et al. 2016). In supervised classification tasks, the Cross-Entropy loss is commonly used to measure the difference between the predicted class probabilities and the true class labels. For supervised regression tasks, the Mean Squared Error loss is typically employed to measure the difference between predicted and true values. In reinforcement learning, the losses often include the TD error or the score function (Sutton and Barto 2018), which converts the expected sum of rewards as a training objective. In

9780192886491_Book.indb 347 27-Mar-25 23:03:45

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

communication games, we often use either a cross-entropy error for the listener or the score function for the speaker. For instance, the cross-entropy would quantify the error of selecting the wrong object in a referential game. In contrast, the score function would quantify how the speaker policy, i.e. emergent language, should be modified according to the collected rewards to solve the task. We explain further these intuitions in section 19.2.3.2.

Optimizing the loss function: The loss function is reduced using a learning process that involves a series of updates known as *Gradient Descent* updates (Rumelhart et al. 1986). They iteratively adjust the network's parameters by following the loss gradient. The magnitude of the update is controlled by a hyperparameter η called the *learning rate*. Given the optimization problem 1, the goal is to find weights such that the loss gradient equals 0. This is achieved by repeating the following gradient update rule:

$$\theta_{n+1} = \theta_n - \eta \nabla_\theta \mathcal{L}_\theta \big|_{\theta_n} \tag{2}$$

where θ_n and θ_{n+1} are the model parameters respectively at iteration n and $n+1$, $\nabla_\theta \mathcal{L}_\theta$ the gradient of the loss function $\mathcal{L}_\theta$ and η the learning rate.

In practice, computing the exact gradient of the averaged loss function $\nabla_\theta \mathcal{L}_\theta$ is infeasible since it necessitates processing the complete dataset. *Stochastic Gradient Descent* (Bottou 2010) overcomes this challenge by approximating the loss function gradient using a limited number of data samples, or *batches*, at each iteration. In standard machine learning libraries (Bradbury et al. 2018; Paszke et al. 2019), *Stochastic Gradient Descent* updates are performed by pre-implemented methods referred to as *optimizers*. In communication games, this gradient is the mathematical operation that modifies the agent behaviour. For instance, every single speaker update alters its generation of symbols, refining its emergent language step after step towards maximizing the reward objective.

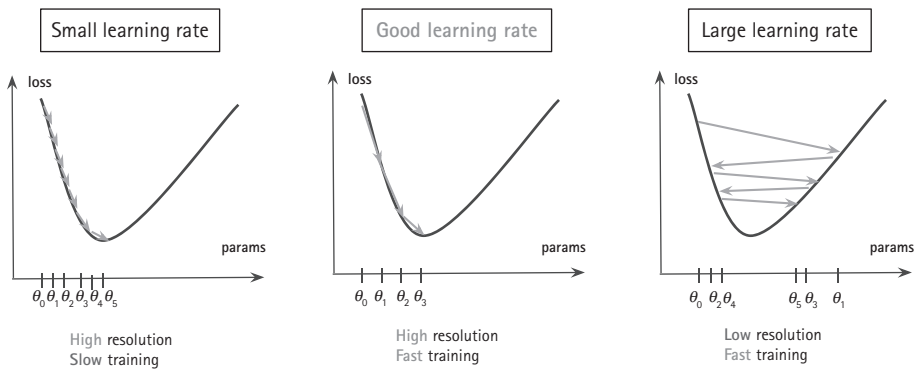


FIGURE 19.9 One-dimensional representation of the effect of adjusting the learning rate.

(a) Small learning rate: parameter updates are too parsimonious; the optimum may be attained, but it requires a large number of updates; (b) Adequate learning rate: the optimum may be reached with precision while limiting the number of updates; (c) Large learning rate: the training may be initially faster but lack the precision to reach the minimum leading to an oscillatory, and sometimes destructive, effect.

9780192886491_Book.indb 348 27-Mar-25 23:03:47

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Good practice: What training parameters to choose, and how to choose them?

1. **Choice of the learning rate** The learning rate controls how strongly a model is adjusted given the loss gradient. As illustrated in Figure 19.9, a too-small learning rate may induce ineffective learning, and a too-large learning rate may cause counterproductive (or even detrimental) (Goodfellow et al. 2016) updates. Fortunately, adaptive optimizer algorithms are designed to tune the learning rate during training. Typically, the learning rate starts large to kickstart the training and is gradually reduced towards the end for fine-tuning. An effective generic choice is to use optimizer *Adam* (Kingma and Ba 2014) optimizer with an initial learning rate of 4.10^{-3}.
2. **Choice of batch size** The batch size controls the gradient update accuracy. Counterintuitively, a large batch size may not guarantee optimal performance. In supervised training, a batch size of sixty-four to 512 samples is often recommended, while in reinforcement learning, larger batch sizes of 256 to $4,096$ samples are preferred.
3. **Balancing batch sizes and learning rates** The batch size and learning rate values are correlated, with no straightforward recipe for interleaving them. It is often worth jointly sweeping over those two hyperparameters to boost performance.

Generalization and overfitting: Training a model involves minimizing the loss of the training data, but evaluating its performance on unseen data is crucial to ensure the network's quality. Intuitively, it is like creating an exam for students with unseen exercises to ensure they correctly understand the lecture. ML Practitioners distinguish (1) the training dataset $\mathcal{D}_{train}$ and its corresponding loss $\mathcal{L}_{train}$, and (2) the test dataset $\mathcal{D}_{test}$ with unseen samples and its corresponding loss $\mathcal{L}_{test}$. The relation between the two losses indicates how well the model generalizes and can be trusted. Figure 19.10 illustrates the three regimes that may occur when training a model:

- **Underfitting**: Both $\mathcal{L}_{train}$ and $\mathcal{L}_{test}$ are high, indicating ineffective learning. An underparametrized network or a small learning rate may cause persistent underfitting. In communication games, this scenario arises when no successful communication emerges between the sender and receiver, resulting in a poor task success both on $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$.
- **Generalization**: Both $\mathcal{L}_{train}$ and $\mathcal{L}_{test}$ are low, indicating successful training and generalization. In communication games, this regime occurs when agents develop a successful communication on $\mathcal{D}_{train}$ that generalizes well to an unseen dataset $\mathcal{D}_{test}$, resulting in high task success both on $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$.
- **Overfitting**: $\mathcal{L}_{train}$ is low, but $\mathcal{L}_{test}$ is high, indicating that the network has recorded the training data and is not able to generalize well to new data. This can be addressed by increasing the amount of training data or using regularization

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

techniques, as explained below. In communication games, this regime occurs when agents develop effective communication on $\mathcal{D}_{train}$ but fail to generalize to an unseen dataset $\mathcal{D}_{test}$.

In communication games, the underfitting regime occurs when the emergent language is not powerful enough to resolve the task, i.e. similar sequences of symbols may represent completely different concepts. Conversely, overfitting occurs when a unique sequence of symbols defines each concept without any structure or compositionally. Therefore, there is no generalization beyond the concepts observed at training time. Optimally, we expect the emergent language to generalize to unseen concepts, which may result from emerging compositionality (Kirby 2001; Rita et al. 2022b).

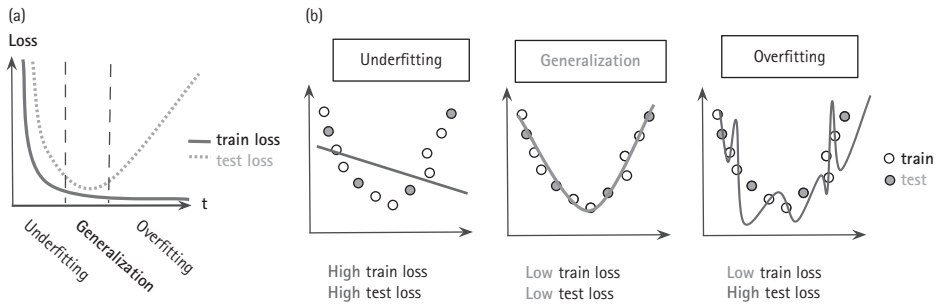


FIGURE 19.10 Training regimes for a one-dimensional regression problem trained with the Mean-Squared Error Loss.

During training, there are three phases: underfitting: both the train and test loss are high, and the corresponding fitting curve does not match nor the training data nor the test data; generalization: both the train and test loss are low, and the fitting curve matches both the training and test data; overfitting: the training loss remains low, but the test loss increases and the corresponding fitting curve perfectly matches the training data but does not generalize to the test data.

Monitoring training: When training a model, it is recommended to divide the dataset into three parts:$\mathcal{D}_{train}$, $\mathcal{D}_{val}$, $\mathcal{D}_{test}$ (typical proportion 80 / 10 / 10). $\mathcal{D}_{train}$ is used to train the model, $\mathcal{D}_{val}$ to find the generalization regime, tune hyperparameters, and retrieve the best model across training, $\mathcal{D}_{test}$ is used to test the model and report the final score. Intuitively, validation data is similar to mock exams, whereas test data is the actual network exam. In practice, the validation loss is regularly plotted and when it starts increasing, training is stopped (Bishop 2006). This technique is known as *Early stopping*.

Regularization methods: Regularization methods were developed to prevent potential overfitting (Goodfellow et al. 2016), as the number of network parameters can be much larger than the data. Some of the most common techniques include:

- **Weight decay**: Overfitting may be caused by excessively increasing parameters. A weight decay penalty can be applied to the training loss. Using the AdamW

9780192886491_Book.indb 350 27-Mar-25 23:03:54

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

variant of the Adam optimizer is recommended to ensure proper integration of the weight penalty (Loshchilov and Hutter 2017).

- **Clipping**: Overfitting may be caused by destructive updates due to unexpected large loss gradients. Clipping methods are applied to cope with such events (Pascanu et al. 2013).
- **Dropout**: Overfitting may be alleviated by only training subsections of networks for each update. This masking mechanism may be applied at the neuron level (Srivastava et al. 2014) or neural-block level for deep networks (Ghiasi et al. 2018).
- **Normalization layers**: High neural activation inside the network tends to deteriorate the training process and favour overfitting. Normalization layers were developed to recalibrate the neural activations, such as batch-normalization (Ioffe and Szegedy 2015), which is a parametrized whitening layer, or layer-normalization (Ba et al. 2016).
- **Data augmentation**: As overfitting often spurs with the lack of data, a common practice is to artificially augment the training set by applying random transformations such as resizing, colour alteration, or partial masking for image data (Ba et al. 2016).

Remark: Applying all regularization techniques simultaneously may seem appealing but can lead to conflicts. For instance, batch normalization should not be applied with dropout, weight decay should not be applied to batch normalization and layer normalization parameters, and strong data augmentations may impair generalization. There is no single rule: finding the right balance for regularization often requires trial and error.

Summary: Figure 19.11 summarizes the training process and the associated experimental choices.

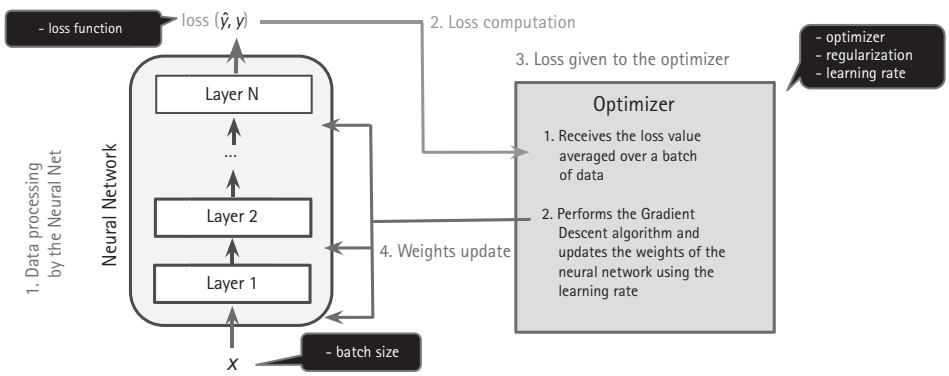


FIGURE 19.11 Updates cycle of a neural network in supervised learning.

First, the neural network receives a batch of input x. The neural network then processes the data and outputs a prediction $\hat{y}$. Second, the loss function computes the average loss value by comparing $\hat{y}$ to the ground truth output y. Third, the average loss value is fed to the optimizer. Last, the optimizer performs a Gradient Descent step to update the neural network weights based on the learning rate. All the experimental choices are framed in speech bubbles.

9780192886491_Book.indb 351 27-Mar-25 23:03:54

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

19.2.3.2 Optimizing communication games with machine learning

Unlike a single network training, two networks are trained simultaneously during a communication game, sometimes requiring different learning methods for each agent. The process involves selecting appropriate (1) learning methods, (2) rewards and loss functions, and (3) optimization protocols.

> **Remark:** The machine learning community has developed frameworks for simulating various communication games, which can be rapidly replicated, understood, and modified. Existing codebases include (Kharitonov et al. 2019) and (Chaabouni et al. 2022) as long as the detailed notebook we provide.

Learning methods: Three learning pipelines are mainly used to train agents in communication games:

1. **Both agents optimized with RL**: This generic and realistic setting assumes no specific task format and involves separate agents with individual rewards and training losses, making it suitable for training any task. However, such training is usually hard to optimize with high variance and requires careful use of RL tools we introduce later.
2. **Sender optimized with RL and Receiver optimized with SL**: This approach is well suited for single-turn message games where the receiver only needs to perform one valid action after receiving a message, such as in referential games (Lewis 1969; Skyrms 2010). In such cases, the receiver's action a is fully determined by the sender's observation x_s and its message m, creating a supervised training sample (m,a) for the receiver. The receiver's training becomes more robust by learning to map messages m to the corresponding output actions a using a supervised loss. Note that the sender still needs to be optimized with RL since message generation is non-differentiable, i.e. the receiver's error cannot propagate to the sender. It ensures more stable training than using a pure RL reward-based approach.
3. **Both agents optimized with SL**: When both agents cooperate fully and optimize the same learning signal, they can be trained using a single supervised training signal. In this scenario, the Sender–Receiver couple is optimized as a single network that maps inputs x_s to output actions a, with a discrete intermediate layer. Reparametrization tricks such as Gumbel-Softmax (Jang et al. 2016; Maddison et al. 2016) have been developed to overcome the non-differentiability of message generation and allow the receiver's error to flow to the sender.[2] Although this approach is more stable than RL methods, we won't go into details because it assumes a less realistic training hypothesis, e.g. the exact error is propagated between sender and receiver as if they were mentally connected.

We next derive the case where agents are optimized with RL as it covers all communication tasks.

[2] This is the same approach as training a Variational Auto-Encoder (VAE) (Kingma and Welling 2013).

9780192886491_Book.indb 352 27-Mar-25 23:03:55

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Reward: Reward functions $\mathcal{R}_s$ and $\mathcal{R}_r$ must be defined to measure the success of the communication task for each agent. These functions typically take agents' observations x_s and x_r and the receiver's action a as input and return 1 if the task is solved, 0 otherwise.

> **Remark:** The reward is the core element inducing the structure of the emergent language. Thus, we recommend carefully avoiding designing rewards towards obtaining a specific language, e.g. directly rewarding compositionality or syntactic properties. Instead, we suggest using rewards that measure communication success without any human prior. Hence, language features may emerge from solving a specific task rather than being forced by design.

The agents' goal is to maximize their respective reward over time, i.e. the expected rewards:

$$\begin{cases} \mathbb{E}_{\tau \sim game\left(\pi_\theta, \rho_\phi\right)}\left[\mathcal{R}_s\right] & \text{Sender's expected reward} \\ \mathbb{E}_{\tau \sim game\left(\pi_\theta, \rho_\phi\right)}\left[\mathcal{R}_r\right] & \text{Receiver's expected reward} \end{cases} \tag{3}$$

$\tau \sim game\left(\pi_\theta, \rho_\phi\right)$ denotes a game episode that depends on the sender's and receiver's stochastic policies. The sender message m and the receiver's action a are sampled from those distributions.

- **Complete formalism**: A complete writing of the expectations should be:

$$\begin{cases} \mathbb{E}_{x_s \sim o_s, x_r \sim o_r, m \sim \pi_\theta(.|x_s), a \sim \rho_\phi(.|m, x_r)}\left[\mathcal{R}_s\left(x_s, x_r, a\right)\right] \\ \mathbb{E}_{x_s \sim o_s, x_r \sim o_r, m \sim \pi_\theta(.|x_s), a \sim \rho_\phi(.|m, x_r)}\left[\mathcal{R}_r\left(x_s, x_r, a\right)\right] \end{cases} \tag{4}$$

where the game is instantiated by sampling the initial agent observations x_s and x_r, the message m is sampled according to the sender's policy $\pi_\theta(.|x_s)$ and the receiver's action a is sampled according to its policy $\rho_\phi(.|m, x_r)$.

- **Stochasticity**: Policy stochasticity is crucial in RL training as it enables the agent to explore its message/action space and learn from its errors. This implies that an object can be described by several messages, and a receiver may take different actions based on a given message.

Loss functions and gradient updates: In reinforcement learning, the goal is to minimize the expected negative reward. However, this objective cannot be directly turned into a gradient update as the reward is not differentiable by definition. Mathematical tools have been developed to circumvent this issue (Sutton and Barto 2018). The policy-gradient algorithm (Sutton et al. 1999) is mostly used in neural language emergence. Denoting by $\nabla_\theta \mathcal{L}_\theta$ and $\nabla_\phi \mathcal{L}_\phi$ the sender and receiver's respective loss gradient, we have:

$$\begin{cases} \nabla_\theta \mathcal{L}_\theta & = -\mathbb{E}_\tau[\nabla_\theta \log \pi_\theta(m|x_s)\mathcal{R}_s] & \text{sender's gradient} \\ \nabla_\phi \mathcal{L}_\phi & = -\mathbb{E}_\tau[\nabla_\phi \log \rho_\phi(a|m, x_r)\mathcal{R}_r] & \text{receiver's gradient} \end{cases} \tag{5}$$

9780192886491_Book.indb 353 27-Mar-25 23:03:59

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

In practice, the quantities (1) $l_\theta = \log \pi_\theta(m \mid x_s) \mid \mathcal{R}_s \mid_{SG}$ and (2) $l_\phi = \log \rho_\phi(a \mid m, x_r) \mid \mathcal{R}_r \mid_{SG}$ are computed over a batch of game episodes and passed to each agent optimizer. $\mid . \mid_{SG}$ is the stop gradient operator that prevents an optimizer from computing the gradient inside the operator.

Optimizing the losses: The optimization encounters challenges, for which we provide a few recipes to ensure a successful optimization process:

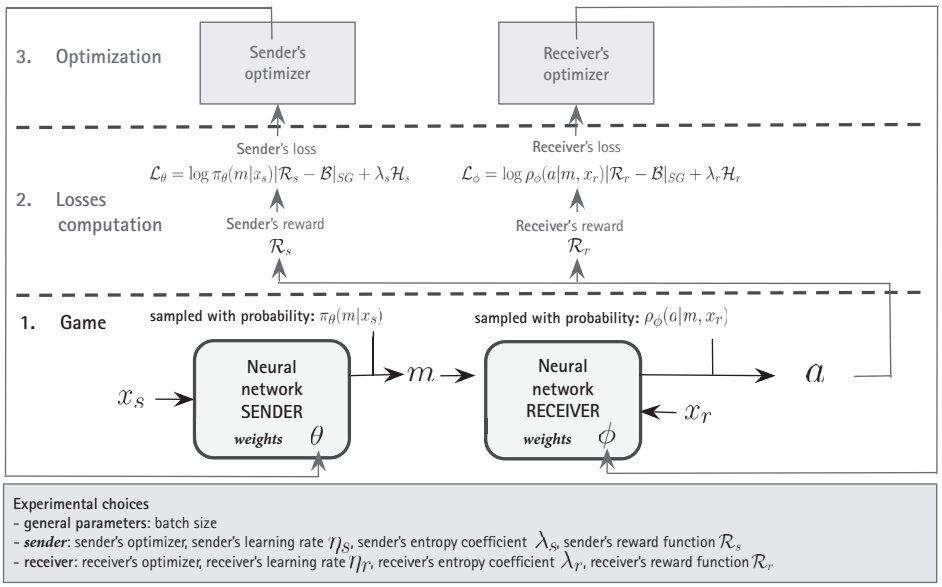


FIGURE 19.12 Scheme of the training loop of a communication game optimized with reinforcement learning.

Step 1 (game): The sender and receiver perform an episode of the game on a batch of data. Agent's observations x_s and x_r, receiver's action a, the probabilities π_θ(m|xs) and $\rho_\phi(a|m, x_r)$ with which the sender and receiver respectively samples the message m and performs action a, the entropy of the sender's policy H_s and receiver's policy H_r are kept to compute the losses. Step 2 (losses computation): Based on those quantities, we compute sender's reward R_s and receiver's reward R_r and then sender's loss L_θ and receiver's loss L_ϕ. Step 3 (optimization): We pass the losses to agents' optimizers that update the weights of the two agents. All the experimental choices are framed below the figure.

1. **Implementing Policy Gradient:** While RL notations may become overwhelming for beginners, their implementation is quite straightforward in practice with recent machine learning libraries (Bradbury et al. 2018; Paszke et al. 2019).
2. **Dealing with large variance**: Estimating the gradient of a RL loss is difficult, due to the large variance of gradient estimates. Large batch sizes and the baseline method should be used to alleviate this. The latter implies subtracting a baseline $\mathcal{B}$ from the reward $\mathcal{R}$, which does not bias the estimate while reducing

9780192886491_Book.indb 354 27-Mar-25 23:04:02

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

the variance. A common baseline is the average value of the reward across a batch of data.

3. **Controlling the exploration–exploitation trade-off**: To prevent the collapse of training due to suboptimal average reward, one can control the exploitation–exploration trade-off by penalizing the entropy of the policies with the terms $\lambda_s \mathcal{H}_s$ and $\lambda_r \mathcal{H}_r$ ($\mathcal{H}_s$ and $\mathcal{H}_r$ refers to the entropy function applied on agents' policies). By increasing the coefficient λ_s(resp. λ_r), the sender's policy (resp. receiver's policy) is encouraged to explore multiple actions instead of focusing on single ones.

As summarized in Figure 19.12, the following optimization protocol can be built applying those practices:

- Choose a batch size and for each agent: learning rates η_s and η_r, reward functions $\mathcal{R}_s$ and $\mathcal{R}_r$, exploration coefficients λ_s and λ_r;
- Iteratively:
  (1) Perform a game episode on a batch of data;
  (2) Compute the losses:

$$\begin{cases} \log \pi_\theta(m \mid x_s) \mid \mathcal{R}_s - \mathcal{B} \mid_{SG} + \lambda_s \mathcal{H}_s & \text{Sender's loss} \\ \log \rho_\phi(a \mid m, x_r) \mid \mathcal{R}_r - \mathcal{B} \mid_{SG} + \lambda_r \mathcal{H}_r & \text{Receiver's loss} \end{cases}$$

  (3) Pass sender's loss (resp. receiver's loss) to sender's optimizer (resp. receiver's optimizer), which performs a parameters update for each agent.

Remark: Training a communication game involves selecting reward functions for each agent and tuning numerous parameters: the batch size, agents' learning rate, and exploration coefficient. The initial step in a simulation is to identify a set of parameters that allows the agents to solve the task. However, it is essential to consider how these choices affect the system's overall training dynamics. Rita et al. (2022a, b) demonstrate that optimization decisions, especially asymmetries between the sender and the receiver, crucially impact the properties of the emergent communication protocol.

19.3 Case study: Simulating a visual discrimination game

We now focus on a particular communication game: the Visual Discrimination Game, a type of Lewis Referential Game (Lewis 1969). These games, which explore how languages emerge through their use, have been extensively studied from theoretical

9780192886491_Book.indb 355 27-Mar-25 23:04:03

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

and experimental angles in language evolution (Crawford and Sobel 1982; Blume et al. 1998; Skyrms 2010; Raviv et al. 2019).

Game rules: The Visual Discrimination Game involves two players: a sender and a receiver. The game proceeds as follows:

- The sender sees an image and communicates about it to the receiver;
- Using the message, the receiver has to guess the original image seen by the sender among a set of N candidate images;
- The original image is revealed, and the two players are informed about the task's success.

Agents play the game repeatedly until they synchronize on a communication protocol that enables the receiver to distinguish any image from any set of distractors.

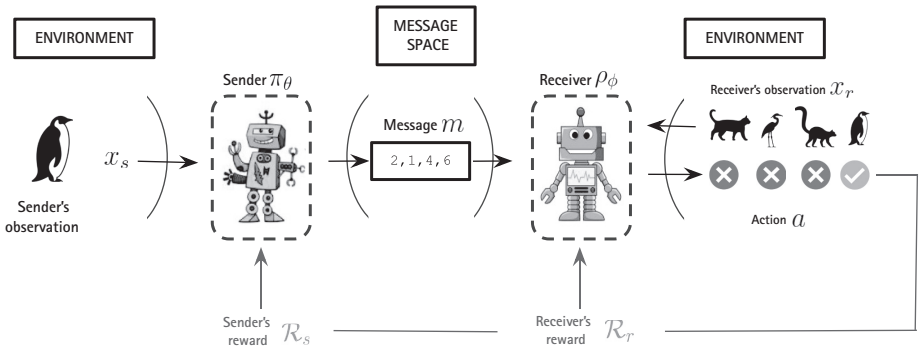


FIGURE 19.13 In the Visual Discrimination Game, the Sender sees an image (sender observation) and communicates about it to the receiver.

Using the message, the receiver has to guess the target image among a set of N candidate images (receiver observation). If the guess (receiver action) is identical to the sender's image, the two agents are equally rewarded.

Designing the game: The following parameters must be specified:

- **Image dataset**: This is the set of images the agents must communicate about. Compared to human simulations, machine learning experiments can be conducted with large-scale datasets compared to human simulations, which is critical for developing a rich communication protocol. For example, some studies, such as Lazaridou et al. (2016); Dessì et al. (2021); Chaabouni et al. (2022); and Rita et al. (2022b), have relied on ImageNet (Deng et al. 2009; Russakovsky et al. 2015) (fourteen million images dataset spanning more than 20,000 categories including animals, vehicles, objects, or instruments). Synthetic datasets, like CLEVR

9780192886491_Book.indb 356 27-Mar-25 23:04:04

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

(Johnson et al. 2017) are also valuable for evaluating agents' ability to communicate about ambiguous images using compositional languages.

- **Number of candidate images N**: The receiver must differentiate the original image from $N-1$ distractor images. The task's difficulty depends on the value of N: a higher N requires a more precise communication protocol.
- **Message space**: The message space is shaped by the vocabulary V and message maximum length L. Adjusting those parameters crucially influences the sender's expressiveness. By denoting the vocabulary size by $|V|$, the sender can use a total number of $L^{|V|}$ messages.

Game formalism: Using previous notations:

- *Sender's observation* x_s is an image sampled from the dataset.
- *Receiver's observation* x_r is a set of N images sampled from the dataset that includes sender's observation x_s.
- *Message* m is a message sent by the sender.
- *Action* a is the choice of image among the set of N images.

Designing the network: Following agents' design of Figure 19.8 and neural networks descriptions in Figure 19.7, Figure 19.14 reports standard agents' design choices in the Visual Discrimination Game.

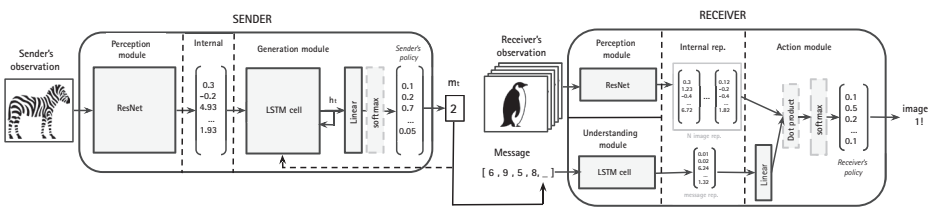


FIGURE 19.14 Examples of neural communicative agents in the Visual Discrimination Game.

Module choices are based on the usual agent design from the literature.

(Left) the sender uses a perception module and a generation module with an internal representation in between. The perception module is a ResNet (He et al. 2016), i.e. CNN with additional skip connections, which produces image representations. The generation module consists of a one-layer LSTM (Hochreiter and Schmidhuber 1997), an RNN variant, followed by a linear layer and a softmax, which generates the sender's policy (probability distribution over the vocabulary symbols). The output is obtained by recursively sampling symbols until the maximum length is reached or the end-of-sentence token is generated.

(Right) the receiver comprises a perception module, an understanding module, and an action module with internal representations inbetween. Like the sender, the perception module is a ResNet, and the understanding module is a one-layer LSTM. The action module includes a linear layer that maps the message representation to a vector with the same dimension as image representations. This dot product compares each image representation to the message representation. Finally, a softmax transforms the resulting N values to probability distributions over the N possible actions. The action, i.e. image choice, is obtained by sampling according to this distribution.

9780192886491_Book.indb 357 27-Mar-25 23:04:06

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Good practice: *Scaling-up neural networks to natural inputs*

- **Pre-trained Network**: Pre-trained modules are employed to simplify computation. These are neural networks that have been previously trained on a different task. In our example, a pre-trained network is used as a perception module, which is frozen during the training (Chaabouni et al. 2022; Rita et al. 2022b). It aims to provide meaningful image representations inherited from the previous task. Chaabouni et al. (2022) provide such pre-trained representations for the ImageNet (Deng et al. 2009; Russakovsky et al. 2015) and CelebA (Liu et al. 2015) datasets.
- **Multimodal Learning**: Fusing different inputs and outputs is referred to as Multimodal Learning (Baltrušaitis et al. 2018). For instance, the receiver needs message and image representations to perform an action. Simple methods like concatenation or scalar product are often used to fuse different neural module representations. When scaling up to more realistic tasks, more advanced fusing mechanisms are sometimes required, e.g. modulation (Dumoulin et al. 2018) or multimodal transformers (Lu et al. 2019; Alayrac et al. 2022).

Optimization: Using Figure 19.12 scheme, a working optimization algorithm using reinforcement learning only is described in Algorithm 1.

Algorithm 1 Visual Discrimination Game optimization

Require: π_θ (sender), ρ_ϕ (receiver) ▷ Trainable neural networks
Require: $\mathcal{D}_{train}$ (image dataset), batch_size, N, V, L ▷ General parameters
Require: sender's optimizer, η_s, $\mathcal{R}$, λ_s, ▷ Sender's parameters
Require: receiver's optimizer, η_r, $\mathcal{R}$, λ_r, ▷ Receiver's parameters

while convergence **do** ▷ **Learning loop**

1. Game

$x_s \sim \mathcal{D}_{train}, x_r \sim \mathcal{D}_{train}$ ▷ Images sampled from the dataset
$\tau \sim \text{game}(\pi_\theta, \rho_\phi, x_s, x_r, V, L)$ ▷ Game episode (batch_size games in parallel)

2. Losses computation

$\text{R} \leftarrow f_{reward}(\tau)$ ▷ Reward computation
$\text{L}_s \leftarrow \log \pi_\theta(m|x_s)|\mathcal{R} - \mathcal{B}|_{SG} + \lambda_s \mathcal{H}_s$ ▷ Sender's loss
$\text{L}_r \leftarrow \log \rho_\phi(a|m, x_r)|\mathcal{R} - \mathcal{B}|_{SG} + \lambda_r \mathcal{H}_r$ ▷ Receiver's loss

3. Optimizer

$\theta \leftarrow$ sender's optimizer$(\theta, \eta_s, \mathcal{L}_s)$ ▷ Sender's weights update
$\phi \leftarrow$ receiver's optimizer$(\theta, \eta_r, \mathcal{L}_r)$ ▷ Receiver's weights update

end while

Parameter choices A typical reward function assigns a reward of 1 if the receiver picks up the correct image and 0 otherwise. The modelling parameters, which include

9780192886491_Book.indb 358 27-Mar-25 23:04:07

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

the vocabulary V, maximum message length L, and the number of candidates N, should be selected based on the problem under investigation. For the optimization, we recommend using a large batch size (512 or 1024 typically) and one Adam (Kingma and Ba 2014) optimizer per agent. The other parameters, including exploration coefficients λ_s, λ_r and learning rates η_s, η_r are interdependent and should be adjusted simultaneously until the simulation works. Common strategies for parameter tuning include manual adjustment or more systematic methods like grid search (Feurer and Hutter 2019).

Implementation: A full implementation of the game with technical details and a starting set of working parameters is provided at: https://github.com/MathieuRita/LangageEvolution_with_DeepLearning.

19.4 Bridging the gap between neural networks and humans in language evolution simulations

This section focuses on current endeavours in using deep learning as a framework for language evolution simulations. It covers the field's progress in using neural networks to replicate human languages and highlights the potential and challenges of deep learning simulations.

19.4.1 Opportunities opened by deep learning simulations

The control/realism duality of simulations: Neural network simulations provide extensive flexibility for modelling various aspects of language emergence simulations, including the game, inputs, and agents. Two primary strategies have been pursued: simplifying experiments into controllable settings (Kottur et al. 2017; Chaabouni et al. 2019, 2020; Ren et al. 2020; Rita et al. 2022a), assessing the influence of incremental modelling elements; and creating more humanly plausible scenarios that emulate language emergence in complex environments (Das et al. 2019; Jaques et al. 2019). It has resulted in various tasks, from basic referential tasks to complex ecological tasks in grounded environments (Das et al. 2019). In terms of inputs, it spans from hand-designed structured and controllable inputs (Kottur et al. 2017; Chaabouni et al. 2019, 2020; Ren et al. 2020; Rita et al. 2020, 2022a) to complicated visual inputs (Evtimova et al. 2017; Lazaridou et al. 2018; Dessì et al. 2021; Chaabouni et al. 2022; Rita et al. 2022b). As for agents, it extends from pairs of agents decomposed into senders and receivers to pairs of bidirectional agents (Bouchacourt and Baroni 2018; Graesser et al. 2019; Taillandier et al. 2023; Michel et al. 2023) and populations (Tieleman et al. 2019; Graesser et al. 2019; Rita et al. 2022a; Michel et al. 2023).

9780192886491_Book.indb 359 27-Mar-25 23:04:08

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Evaluating emergent linguistic phenomena: Simulations give rise to the emergence of artificial languages whose properties are compared to human languages. As human languages can be described in terms of language universals, i.e. abstract properties found across all human languages, studies have tried to establish the conditions under which those universal properties emerge. Such universals mainly include compositionality, i.e. the ability to decompose the meaning of an utterance as a function of its constituents (Hockett 1960), measured through topographic similarity (Brighton and Kirby 2006; Chaabouni et al. 2020), or Tree Reconstruction Error (Andreas 2019); efficiency, i.e. efficient information compression, measured through message length statistics and semantic categorization (Zipf 1949; Regier et al. 2015); and demographic trends, such as the impact of population size, contact agents proportion, network topology on language structure (Clyne 1992; Wray and Grace 2007; Wagner 2009; Lupyan 2010).

19.4.2 Do neural networks replicate human behaviours?

To provide valuable insights through deep learning simulations, replicating human languages is essential. This involves identifying the basic assumptions needed for artificial agents to display human-like language patterns in their communication protocols.

The referential objective is insufficient for the emergence of natural language features: A first approach is to question whether the most simple communication task, i.e. referring to objects in an environment through referential communication, is enough to see human language features emerge. The first works on referential tasks showed that neural agents could successfully derive a communication protocol from solving the task (Lazaridou et al. 2016; Kottur et al. 2017; Havrylov and Titov 2017). Still, such protocols are neither interpretable nor bear the core properties of human languages. Indeed, agents tasked with communicating about images do not utilize semantically significant concepts but instead shortcut the task by basing their messages on low-level visual features (Lazaridou et al. 2016; Havrylov and Titov 2017; Bouchacourt and Baroni 2018; Chaabouni et al. 2022). Additionally, when agents communicate about hand-designed structured sets of objects in a simple referential task, fundamental properties of natural languages such as compositionality (Kottur et al. 2017; Chaabouni et al. 2020) or efficiency (Zipf 2016) do not spontaneously arise (Chaabouni et al. 2019). Eventually, when referential games are played within a population of agents, human demographic trends are not reproduced. Population size does not behave as a regularization factor (Li and Bowling 2019; Cogswell et al. 2019; Rita et al. 2022a; Chaabouni et al. 2022) and agents do not synchronize on a shared protocol (Rita et al. 2022a; Michel et al. 2023). Understanding the origins of these discrepancies from either an optimization or modelization perspective is an active research question.

Incorporating human-inspired constraints drive the emergence of natural language features: To recover human languages features, different human-inspired constraints have incrementally been added to simulations. Inspired by Iterated Learning (Kirby 2001; Kirby et al. 2014), one line of research has explored the effects of

9780192886491_Book.indb 360 27-Mar-25 23:04:08

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

learnability constraints on language emergence by altering learning dynamics. Li and Bowling (2019), Ren et al. (2020), and Cogswell et al. (2019) implement neural variants of Iterated Learning by periodically introducing newborn agents and mimicking generational transmission. They find that those learning constraints drive the selection of more compositional languages, as they are easier to learn (Li and Bowling 2019). Another line of research focuses on incorporating cognitively inspired biases into agent modelling. For example, Rita et al. (2020) show that Zipf's Law of Abbreviation (Zipf 1949) emerges when both pressures towards least-effort production (Zipf 1949; Piantadosi et al. 2011; Kanwal et al. 2017) and comprehension laziness are introduced. Eventually, some researchers have refined population modelling. Rita et al. (2022a) introduce learning speed variations into populations and recover the relationship between population size and language structure reported in previous works (Lupyan 2010; Meir et al. 2012; Dryer and Haspelmath 2013; Reali et al. 2018; Raviv et al. 2019). Graesser et al. (2019) examine contact agents phenomena and show that a contact language can either converge towards the majority protocol or result in novel creole languages, depending on the inter- and intra-community densities. Kim and Oh (2021) and Michel et al. (2023) study how social graph connectivity impacts the development of shared languages.

19.4.3 Towards realistic experiments

Although incorporating human-inspired constraints shows promise for replicating human language features, the simplicity of current models remains limited. An avenue is opened for the design of humanly plausible experiments. We present efforts to build more realistic models and discuss the associated challenges here.

19.4.3.1 Towards realistic scenarios

Task-specific communication games may be restrictive as they overlook other aspects of our language, such as conversation, interaction with the physical world, and other modalities. More realistic scenarios are needed to encompass all aspects of our language. Some attempts have been made to create more plausible settings. Chaabouni et al. (2022) complexify the referential task by scaling the game to large datasets and tasking agents to retrieve images among 1,000 distractors. Evtimova et al. (2017) and Taillandier et al. (2023) model conversation by building bidirectional agents for multiturn communications; Bullard et al. (2020) explore non-verbal communication using spatially articulated agents; Das et al. (2019) ground agents in more realistic two-dimensional (2D) and three-dimensional (3D) environments; Jaques et al. (2019) test agents' ability to solve social dilemmas in grounded environments. However, making more realistic games poses both technical and analytical challenges. Training instabilities can occur when games become more complex, requiring optimization tricks (Chaabouni et al. 2022). Moreover, as environments become more complex, the emergence of language is more challenging to analyse. For example, Lowe et al.

9780192886491_Book.indb 361 27-Mar-25 23:04:08

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

(2020) demonstrate how agents can solve complex tasks with shallow communication protocols and why new tools are needed to assess emergent languages qualitatively and quantitatively in these situations.

19.4.3.2 Towards realistic agents

Many neural communication agents are designed for specific games and lack crucial aspects of human cognition. For instance, agents are often limited to either speaking or listening, which overlooks the interplay between comprehension and production (Galke et al. 2022). Some works propose more realistic agents. These include bidirectional agents that both speak and listen (Bouchacourt and Baroni 2018; Graesser et al. 2019; Michel et al. 2023; Taillandier et al. 2023), as well as agents with restricted memory capacity that better mirrors human cognition (Resnick et al. 2019). Additionally, Rita et al. (2020) incorporate the least-effort principle to make agents efficient encoders (Zipf 1949; Piantadosi et al. 2011; Kanwal et al. 2017). Still, despite the impact of these modelling constraints on emergent language properties, they are not consistently applied across the literature.

19.4.3.3 Towards linguistically informed metrics

One of the main limitations of neural emergent languages is that current metrics may not capture crucial features of human languages. For instance, most work only uses topographic similarity (Lazaridou et al. 2018; Li and Bowling 2019) as a structural metric (Brighton and Kirby 2006), which assumes that the units of the message carry out the meaning. In human languages, the meaning units are the results of a combinatorial process using non-meaningful units, such as phonetic features or phonemes (the so-called double articulation phenomenon (Martinet 1960), or duality of patterning (Hockett 1970)). Other universal properties of language (formal universals; Chomsky and Halle 1968) include the reliance on symbols and rules (Fodor and Pylyshyn 1988), the use of hierarchical representations or long-distance dependencies (Hauser et al. 2002), the existence of part-of-speech classes (Rijkhoff 2007), such as the distinction between content and grammatical words, the existence of deixis (Lyons 1977), i.e. the use of certain parts of the message to refer to places or time or person relative to the context of elocution of the message, and many others. Studying such properties is challenging as it requires the design of adapted measures that could be computed both on human and artificial languages. Furthermore, current artificial settings are often too simple to drive the emergence of such properties, reinforcing the need for more realistic scenarios that translate into our environment's complexity.

19.4.3.4 Bridging natural language processing and language emergence

One area of research focuses on investigating whether language emergence simulations can potentially enhance natural language processing tasks. One approach involves pre-training language models with artificial languages that emerged from communication games, resulting in a moderate boost when fine-tuning low-resource language tasks (Yao et al. 2022). Another approach is exploring machine–machine interaction to learn an emergent communication protocol that prompts large language models (Shin

9780192886491_Book.indb 362 27-Mar-25 23:04:08

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

et al. 2020; Deng et al. 2022). Reciprocally, natural language models can be utilized to explore language evolution from pre-trained languages, such as studying creolization (Armstrong et al. 2022) or language drift phenomena (Lu et al. 2020). Finally, at the time of writing, Large Language Models (LLM) (Bai et al. 2022; Ouyang et al. 2022; Hoffmann et al. 2022b; Touvron et al. 2023) have demonstrated potential for Natural Language Processing, as they can handle multiple languages and perform basic reasoning. This presents exciting opportunities for language emergence research from scientific and practical perspectives (Baroni et al. 2022).

19.5 Conclusion

Deep learning advancements offer new opportunities for simulating language evolution, as neural networks can handle diverse data without pre-defined human priors. They scale significantly regarding dataset size, task complexity, and number of participants or generations. This opens up possibilities for creating realistic language evolution scenarios at unprecedented scales. Reciprocally, language evolution research can provide valuable insights for developing future deep learning models. In the journey towards building intelligent language models, it seems essential to incorporate constraints and mechanisms that shape the development and evolution of language, such as perceptual, social, or environmental pressures. We hope this chapter will encourage researchers in both language evolution and deep learning to collaborate and jointly explore those two captivating blackboxes: humans and neural networks.

References

Alayrac, J.-B., Donahue, J., Luc, P., et al. (2022). 'Flamingo: A visual language model for few-shot learning', *Advances in Neural Information Processing Systems* 35: 23716–23736.

Andreas, J. (2019). 'Measuring compositionality in representation learning', *arXiv preprint arXiv:1902.07181*.

Armstrong, R.-A., Hewitt, J., and Manning, C. (2022). 'Jampatoisnli: A Jamaican patois natural language inference dataset', *arXiv preprint arXiv:2212.03419*.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). 'Neural machine translation by jointly learning to align and translate', *arXiv preprint arXiv:1409.0473*.

Bai, Y., Kadavath, S., Kundu, S., et al. (2022). 'Constitutional AI: Harmlessness from AI feedback', *arXiv preprint arXiv:2212.08073*.

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. (2018). 'Multimodal machine learning: A survey and taxonomy', *IEEE Transactions on Pattern Analysis and Machine Intelligence* 410(2): 423–443.

Baronchelli, A., Gong, T., Puglisi, A., and Loreto, V. (2010). 'Modeling the emergence of universality in color naming patterns', *Proceedings of the National Academy of Sciences* 1070(6): 2403–2407.

9780192886491_Book.indb 363 27-Mar-25 23:04:08

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Baroni, M., Dess, R., and Lazaridou, A. (2022). 'Emergent language-based coordination in deep multi-agent systems', in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, 11–16.

Bickerton, D. (2007). 'Language evolution: A brief guide for linguists', *Lingua* 1170(3): 510–526.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Berlin, Heidelberg: Springer-Verlag.

Blume, A., DeJong, D. V., Kim, Y.-G., and Sprinkle, G. B. (1998). 'Experimental evidence on the evolution of meaning of messages in sender-receiver games', *The American Economic Review* 880(5): 1323–1340.

Bottou, L. (2010). 'Large-scale machine learning with stochastic gradient descent', in *Proceedings of COMPSTAT'2010: 19th International Conference on Computational Statistics, Paris France, August 22–27, 2010 Keynote, Invited and Contributed Papers*, 177–186. Berlin: Springer.

Bouchacourt, D. and Baroni, M. (2018). 'How agents see things: On visual representations in an emergent language game', in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 981–985, Brussels: Association for Computational Linguistics, https://aclanthology.org/D18-1119.

Brace, L., Bullock, S., and Noble, J. (2015). 'Achieving compositional language in a population of iterated learners', in *ECAL 2015: The 13th European Conference on Artificial Life*, 349–356. Cambridge, MA: MIT Press.

Bradbury, J., Frostig, R., Hawkins, P., et al. (2018). 'JAX: Composable transformations of Python+NumPy programs', http://github.com/google/jax.

Brighton, H. and Kirby, S. (2006). 'Understanding linguistic evolution by visualizing the emergence of topographic mappings', *Artificial Life* 120(2): 229–242.

Brown, T., Mann, B., Ryder, N., et al. (2020). 'Language models are few-shot learners', *Advances in Neural Information Processing Systems* 33: 1877–1901.

Bullard, K., Meier, F., Kiela, D., et al. (2020). 'Exploring zero-shot emergent communication in embodied multi-agent populations', *arXiv preprint arXiv:2010.15896*.

Carroll, M., Shah, R., Ho, M. K., et al. (2019). 'On the utility of learning about humans for human-AI coordination', *Advances in Neural Information Processing Systems* 32.

Chaabouni, R., Kharitonov, E., Dupoux, E., and Baroni, M. (2019). 'Anti-efficient encoding in emergent communication', *Advances in Neural Information Processing Systems* 32.

Chaabouni, R., Kharitonov, E., Bouchacourt, D., et al. (2020). 'Compositionality and generalization in emergent languages', *arXiv preprint arXiv:2004.09124*.

Chaabouni, R., Strub, F., Altché, F., et al. (2022). 'Emergent communication at scale', in *Proceedings of International Conference on Learning Representations (ICLR)*.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). 'A simple framework for contrastive learning of visual representations', in *International Conference on Machine Learning* 1597–1607. PMLR.

Cho, I.-K. and Kreps, D. M. (1987). 'Signaling games and stable equilibria*', *The Quarterly Journal of Economics* 1020(2): 179–221, ISSN 0033-5533, https://doi.org/10.2307/1885060.

Chomsky, N. and Halle, M. (1968). *The Sound Pattern of English*. ERIC.

Chung, J., Gulcehre, C., Cho, K.-H., and Bengio, Y. (2014). 'Empirical evaluation of gated recurrent neural networks on sequence modeling', *arXiv preprint arXiv:1412.3555*.

Clyne, M, (1992). 'Linguistic and sociolinguistic aspects of language contact, maintenance and loss', *Maintenance and Loss of Minority Languages* 1: 17.

9780192886491_Book.indb 364 27-Mar-25 23:04:08

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Cogswell, M., Lu, J., Lee, S., et al. (2019). 'Emergence of compositional language with deep generational transmission', *arXiv preprint arXiv:1904.09067*.

Crawford, V. P. and Sobel, J. (1982). 'Strategic information transmission', *Econometrica: Journal of the Econometric Society* 50: 1431–1451.

Cuskley, C., Castellano, C., Colaiori, F., et al. (2017). 'The regularity game: Investigating linguistic rule dynamics in a population of interacting agents', *Cognition* 159: 25–32.

Das, A., Gervet, T., Romoff, J., et al. (2019). 'Tarmac: Targeted multi-agent communication', in *International Conference on Machine Learning*, 1538–1546. PMLR.

Deng, J., Dong, W., Socher, R., et al. (2009). 'Imagenet: A large-scale hierarchical image database', in *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*.

Deng, M., Wang, J., Hsieh, C.-P., et al. (2022). 'RLprompt: Optimizing discrete text prompts with reinforcement learning', *arXiv preprint arXiv:2205.12548*.

Dess, R., Kharitonov, E., and Marco, B. (2021). 'Interpretable agent communication from scratch (with a generic visual processor emerging on the side)', *Advances in Neural Information Processing Systems* 34: 26937–26949.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). 'Bert: Pre-training of deep bidirectional transformers for language understanding', *arXiv preprint arXiv:1810.04805*.

Dryer, M. S. and Haspelmath, M. (eds.). (2013). *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, https://wals.info/.

Dumoulin, V., Perez, E., Schucher, N., et al. (2018). 'Feature-wise transformations', *Distill* 30(7): e11.

Elman, J. L. (1990). 'Finding structure in time', *Cognitive Science* 140(2): 179–211.

Evtimova, K., Drozdov, A., Kiela, D., and Cho, K. (2017). 'Emergent communication in a multi-modal, multi-step referential game', *arXiv preprint arXiv:1705.10369*.

Feurer, M. and Hutter, F. (2019). *Hyperparameter Optimization*. Berlin, Heidelberg: Springer International Publishing, 3–33.

Fodor, J. A. and Pylyshyn, Z. W. (1988). 'Connectionism and cognitive architecture: A critical analysis', *Cognition* 280(1–2): 3–71.

Fukushima, K. (1980). 'Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position', *Biological Cybernetics* 360(4): 193–202.

Galke, L., Ram, Y., and Raviv, L. (2022). 'Emergent communication for understanding human language evolution: What's missing?', https://arxiv.org/abs/2204.10590.

Ghiasi, G., Lin, T.-Y., and Le, Q. V. (2018). 'Dropblock: A regularization method for convolutional networks', *Advances in Neural Information Processing Systems* 31.

Gong, T., Shuai, L., and Zhang, M. (2014). 'Modelling language evolution: Examples and predictions', *Physics of Life Reviews* 110(2): 280–302.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Cambridge, MA: MIT Press.

Graesser, L., Cho, K., and Kiela, D. (2019). 'Emergent linguistic phenomena in multi-agent communication games', *arXiv preprint arXiv:1901.08706*.

Harnad, S., Steklis, H. D., and Lancaster, J. B. (1976). 'Origins and evolution of language and speech', *Language* 540(3): 647–660.

Harsanyi, J. C. (1967). 'Games with incomplete information played by "Bayesian" players, i–iii part i. the basic model', *Management Science* 140(3): 159–182.

Hauser, M. D. (1996). *The Evolution of Communication*. Cambridge, MA: MIT Press.

9780192886491_Book.indb 365 27-Mar-25 23:04:08

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Hauser, M. D., Chomsky, N., and Tecumseh Fitch, W. (2002). 'The faculty of language: What is it, who has it, and how did it evolve?', *Science* 2980(5598): 1569–1579.

Havrylov, S. and Titov, I. (2017). 'Emergence of language with multi-agent games: Learning to communicate with sequences of symbols', *Advances in Neural Information Processing Systems* 30. Cambridge, MA: MIT Press.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). 'Deep residual learning for image recognition', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hochreiter, S. and Schmidhuber, J. (1997). 'Long short-term memory', *Neural Computation* 90(8): 1735–1780.

Hockett, C. F. (1960). 'The origin of speech', *Scientific American* 2030(3): 88–97.

Hockett, C. F. (1970). *A Leonard Bloomfield Anthology*. ERIC.

Hoffmann, J., Borgeaud, S., Mensch, A., et al. (2022a). 'An empirical analysis of compute-optimal large language model training', *Advances in Neural Information Processing Systems* 35: 30016–30030.

Hoffmann, J., Borgeaud, S., Mensch, A., et al. (2022b). 'Training compute-optimal large language models', *arXiv preprint arXiv:2203.15556*.

Hornik, K., Stinchcombe, M., and White, H. (1989). 'Multilayer feedforward networks are universal approximators', *Neural Networks* 20(5): 359–366.

Hurford, J. R. (1989). 'Biological evolution of the Saussurean sign as a component of the language acquisition device', *Lingua* 770(2): 187–222.

Ioffe, S. and Szegedy, C. (2015). 'Batch normalization: Accelerating deep network training by reducing internal covariate shift', in *International Conference on Machine Learning*, 448–456. PMLR.

Jang, E., Gu, S., and Poole, B. (2016). 'Categorical reparameterization with Gumbel-Softmax', *arXiv preprint arXiv:1611.01144*.

Jaques, N., Lazaridou, A., Hughes, E., et al. (2019). 'Social influence as intrinsic motivation for multi-agent deep reinforcement learning', in *International Conference on Machine Learning*, 3040–3049. PMLR.

Johnson, J., Hariharan, B., Van Der Maaten, L., et al. (2017). 'Clevr: A diagnostic dataset for compositional language and elementary visual reasoning', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2901–2910.

Kalinowska, A., Davoodi, E., Strub, F., et al. (2022). 'Over-communicate no more: Situated RL agents learn concise communication protocols', *arXiv preprint arXiv:2211.01480*.

Kanwal, J., Smith, K., Culbertson, J., and Kirby, S. (2017). 'Zipf's law of abbreviation and the principle of least effort: Language users optimise a miniature lexicon for efficient communication', *Cognition* 165: 45–52.

Ke, J., Gong, T., Wang, W. S. Y., et al. (2008). 'Language change and social networks', *Communications in Computational Physics* 30(4): 935–949.

Kharitonov, E., Chaabouni, R., Bouchacourt, D., and Baroni. M. (2019). 'Egg: A toolkit for research on emergence of language in games', *arXiv preprint arXiv:1907.00852*.

Kim, J. and Oh, A. (2021). 'Emergent communication under varying sizes and connectivities', *Advances in Neural Information Processing Systems* 34: 17579–17591.

Kingma, D. P. and Ba, J. (2014). 'Adam: A method for stochastic optimization', *arXiv preprint arXiv:1412.6980*.

Kingma, D. P. and Welling, M. (2013). 'Auto-encoding variational Bayes', *arXiv preprint arXiv:1312.6114*.

9780192886491_Book.indb 366 27-Mar-25 23:04:08

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Kipf, T. N. and Welling, M. (2016). 'Semi-supervised classification with graph convolutional networks', *arXiv preprint arXiv:1609.02907*.

Kirby, S. (2001). 'Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity', *IEEE Transactions on Evolutionary Computation* 50(2): 102–110.

Kirby, S. (2002). 'Natural language from artificial life', *Artificial Life* 80(2): 185–215.

Kirby, S., Cornish, H., and Smith, K. (2008). 'Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language', *Proceedings of the National Academy of Sciences* 1050(31): 10681–10686.

Kirby, S., Griffiths, T., and Smith, K. (2014). 'Iterated learning and the evolution of language', *Current Opinion in Neurobiology* 28: 108–114.

Kirby, S., Tamariz, M., Cornish, H., and Smith, K. (2015). 'Compression and communication in the cultural evolution of linguistic structure', *Cognition* 141: 87–102.

Kottur, S., Moura, J. M. F., Lee, S., and Batra, D. (2017). 'Natural language does not emerge "naturally" in multi-agent dialog', *arXiv preprint arXiv:1706.08502*.

Lazaridou, A. and Baroni, M. (2020). 'Emergent multi-agent communication in the deep learning era', *arXiv preprint arXiv:2006.02419*.

Lazaridou, A., Peysakhovich, A., and Baroni, M. (2016). 'Multi-agent cooperation and the emergence of (natural) language', *arXiv preprint arXiv:1612.07182*.

Lazaridou, A., Hermann, K. M., Tuyls, K., and Clark, S. (2018). 'Emergence of linguistic communication from referential games with symbolic and pixel input', *arXiv preprint arXiv:1804.03984*.

LeCun, Y., Touresky, D., Hinton, G., and Sejnowski, T. (1988). 'A theoretical framework for back-propagation', in *Proceedings of the 1988 Connectionist Models Summer School* 1: 21–28.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). 'Gradient-based learning applied to document recognition', *Proceedings of the IEEE* 860(11): 2278–2324.

Lei Ba, J., Kiros, J. R., and Hinton, G. E. (2016). 'Layer normalization', *arXiv preprint arXiv:1607.06450*.

Lewis, D. K. (1969). *Convention: A Philosophical Study*. Cambridge, MA: Wiley-Blackwell.

Li, F. and Bowling, M. (2019). 'Ease-of-teaching and language structure from emergent communication', in *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*.

Lieberman, P. and Crelin, E. S. (1971). 'On the speech of Neanderthal man', *JANUA LINGUARUM* 76.

Littman, M. L. (1994). 'Markov games as a framework for multi-agent reinforcement learning', in *Machine Learning Proceedings 1994*, 157–163. Amsterdam: Elsevier.

Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). 'Deep learning face attributes in the wild', in *Proceedings of the International Conference on Computer Vision (ICCV)*.

Loshchilov, I. and Hutter, F. (2017). 'Decoupled weight decay regularization', *arXiv preprint arXiv:1711.05101*.

Lowe, R., Gupta, A., Foerster, J., et al. (2020). 'On the interaction between supervision and self-play in emergent communication', *arXiv preprint arXiv:2002.01093*.

Lu, J., Batra, D., Parikh, D., and Lee, S. (2019). 'Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks', *Advances in Neural Information Processing Systems* 32.

Lu, Y., Singhal, S., Strub, F., et al. (2020). 'Countering language drift with seeded iterated learning', in *International Conference on Machine Learning*, 6437–6447. PMLR.

9780192886491_Book.indb 367 27-Mar-25 23:04:08

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Lupyan, R. D. G. (2010). 'Language structure is partly determined by social structure', *PLoS ONE* 5: 1.

Lyons, J. (1977). *Semantics: Volume 2*, Cambridge: Cambridge University Press.

Maddison, C. J., Mnih, A., and Whye Teh, Y. (2016). 'The concrete distribution: A continuous relaxation of discrete random variables', *arXiv preprint arXiv:1611.00712*.

Martinet, A. (1960). 'Elements of a functional syntax', *Word* 160(1):1–10.

Meir, I., Israel, A., Sandler, W., et al. (2012). 'The influence of community on language structure: Evidence from two young sign languages', *Linguistic Variation* 120(2): 247–291.

Michel, P., Rita, M., Mathewson, K. W., et al. (2023). 'Revisiting populations in multi-agent communication', [under review].

Mikolov, T., Karafiát, M., Burget, L., et al. (2010). 'Recurrent neural network based language model', in *Interspeech*, 1045–1048.

Mitchell, T. M. (1997). *Machine Learning*. New York: McGraw-Hill, Inc, 1 edition.

Motamedi, Y., Schouwstra, M., Smith, K., et al. (2019). 'Evolving artificial sign languages in the lab: From improvised gesture to systematic sign', *Cognition* 192: 103964.

Myers-Scotton, C. (2002). *Contact Linguistics: Bilingual Encounters and Grammatical Outcomes*. Oxford: Oxford University Press.

Nair, V. and Hinton, G. E. (2010). 'Rectified linear units improve restricted Boltzmann machines', in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 807–814.

Ouyang, L., Wu, J., Jiang, X., et al. (2022). 'Training language models to follow instructions with human feedback', *Advances in Neural Information Processing Systems* 35: 27730–27744.

Pascanu, R., Mikolov, T., and Bengio, Y. (2013). 'On the difficulty of training recurrent neural networks', in *International Conference on Machine Learning*, 1310–1318. PMLR.

Paszke, A., Gross, S., Chintala, S., et al. (2017). 'Automatic differentiation in PyTorch', in *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*.

Paszke, A., Gross, S., Massa, F., et al. (2019). 'PyTorch: An imperative style, high-performance deep learning library', *Advances in Neural Information Processing Systems* 32.

Piantadosi, S. T., Tily, H., and Gibson, E. (2011). 'Word lengths are optimized for efficient communication', *Proceedings of the National Academy of Sciences* 1080(9): 3526–3529.

Raviv, L., Meyer, A., and Lev-Ari, S. (2019). 'Larger communities create more systematic languages', *Proceedings of the Royal Society B* 2860(1907): 20191262.

Reali, F. and Griffiths, T. L. (2009). 'The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning', *Cognition* 1110(3): 317–328.

Reali, F. and Griffiths, T. L. (2010). 'Words as alleles: Connecting language evolution with Bayesian learners to models of genetic drift', *Proceedings of the Royal Society B: Biological Sciences* 2770(1680): 429–436.

Reali, F., Chater, N., and Christiansen, M. H. (2018). 'Simpler grammar, larger vocabulary: How population size affects language', *Proc. of the Royal Society B: Biological Sciences* 2850(1871): 20172586. doi: 10.1098/rspb.2017.2586.

Regier, T., Kemp, C., and Kay, P. (2015). 'Word meanings across languages support efficient communication', in B. MacWhinney and W. O'Grady (eds.), *The Handbook of Language Emergence*, 237–263. New York: Wiley-Blackwell.

Ren, Y., Guo, S., Labeau, M., et al. (2020). 'Compositional languages emerge in a neural iterated learning model', *arXiv preprint arXiv:2002.01365*.

Resnick, C., Gupta, A., Foerster, J., et al. (2019). 'Capacity, bandwidth, and compositionality in emergent language learning', *arXiv preprint arXiv:1910.11424*.

9780192886491_Book.indb 368 27-Mar-25 23:04:08

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Rijkhoff, J. (2007). 'Word classes', *Language and Linguistics Compass* 10(6): 709–726.
Rita, M., Chaabouni, R., and Dupoux, E. (2020). '"LazImpa": Lazy and impatient neural agents learn to communicate efficiently', *arXiv preprint arXiv:2010.01878*.
Rita, M., Strub, F, Grill, J.-B., et al. (2022a). 'On the role of population heterogeneity in emergent communication', in *Proceedings of International Conference on Learning Representations (ICLR)*.
Rita, M., Tallec, C., Michel, P., et al. (2022b). 'Emergent communication: Generalization and overfitting in Lewis games', *arXiv preprint arXiv:2209.15342*.
Rosenblatt, F. (1958). 'The perceptron: A probabilistic model for information storage and organization in the brain', *Psychological Review* 650(6): 386.
Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). 'Learning representations by back-propagating errors', *Nature* 3230(6088): 533–536.
Russakovsky, O., Deng, J., Su, H., et al. (2015). 'Imagenet large scale visual recognition challenge', *International Journal of Computer Vision (IJCV)* 1150(3): 211–252.
Shannon, C. E. (1948). 'A mathematical theory of communication', *The Bell System Technical Journal* 270(3): 379–423.
Shin, T., Razeghi, Y., Logan IV, R. L., et al. (2020). 'Autoprompt: Eliciting knowledge from language models with automatically generated prompts', *arXiv preprint arXiv:2010.15980*.
Silver, D., Huang, A., Maddison, C. J., et al. (2016). 'Mastering the game of go with deep neural networks and tree search', *Nature* 5290(7587): 484–489.
Silver, D., Hubert, T., Schrittwieser, J., et al. (2017). 'Mastering chess and shogi by self-play with a general reinforcement learning algorithm', *arXiv preprint arXiv:1712.01815*.
Skyrms, B. (2010). *Signals: Evolution, Learning, and Information*. Oxford: Oxford University Press.
Smith, K., Brighton, H., and Kirby, S. (2003). 'Complex systems in language evolution: The cultural emergence of compositional structure', *Advances in Complex Systems* 60(4): 537–558.
Srivastava, N., Hinton, G., Krizhevsky, A., et al. (2014). 'Dropout: A simple way to prevent neural networks from overfitting', *The Journal of Machine Learning Research* 150(1): 1929–1958.
Steels, L. (1995). 'A self-organizing spatial vocabulary', *Artificial Life* 20(3): 319–332.
Steels, L. (1997). 'The synthetic modeling of language origins', *Evolution of Communication* 10(1): 1–34.
Sutskever, I., Martens, J., and Hinton, G. E. (2011). 'Generating text with recurrent neural networks', in *ICML*.
Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.
Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. (1999). 'Policy gradient methods for reinforcement learning with function approximation', *Advances in Neural Information Processing Systems* 12.
Tadelis, S. (2013). *Game Theory: An Introduction*. Princeton, NJ: Princeton University Press.
Taillandier, V., Hupkes, D., Sagot, B., et al. (2023). 'Neural agents struggle to take turns in bidirectional emergent communication', in *The Eleventh International Conference on Learning Representations*.
Tieleman, O., Lazaridou, A., Mourad, S., et al. (2019). 'Shaping representations through communication: Community size effect in artificial learning system', in *Visually Grounded Interaction and Language (ViGIL) Workshop*.
Touvron, H., Lavril, T., Izacard, G., et al. (2023). 'Llama: Open and efficient foundation language models', *arXiv preprint arXiv:2302.13971*.

9780192886491_Book.indb 369 27-Mar-25 23:04:08

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). 'Attention is all you need', *Advances in Neural Information Processing Systems* 30.

Vogt, P. (2009). 'Modeling interactions between language evolution and demography', *Human Biology* 810(3): 237–258.

Wagner, E. (2009). 'Communication and structured correlation', *Erkenntnis* 710(3): 377–393.

Wray, A. and Grace, G. W. (2007). 'The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form', *Lingua* 1170(3): 543–578. ISSN 0024-3841. The Evolution of Language.

Yao, S., Yu, M., Zhang, Y., et al. (2022). 'Linking emergent and natural languages via corpus transfer', *arXiv preprint arXiv:2203.13344*.

Zipf, G. K. (1949). *The Principle of Least Effort*. New York: Hafner Publishing Company.

Zipf, G. K. (2016). *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. New York: Ravenio Books.

9780192886491_Book.indb 370 27-Mar-25 23:04:08

CHAPTER 20

TASK-DRIVEN LANGUAGE EVOLUTION WITH SWARM ROBOTICS

NICOLAS CAMBIER AND ROMAN MILETITCH

20.1 Introduction

Language games, extensively studied within multi-agent systems, have seen a growing adoption in swarm robotics (Cambier et al. 2020). Swarm robotics takes inspiration from social insects to model and study behaviours of large numbers of simple embodied robots tackling together a common task (Brambilla et al. 2013). The aim of such studies is to show self-organization at the level of the swarm: the emergence of a collective behaviour that is more complex than the sum of isolated individual behaviour. To reach that end, an experimentation is the sum of: (a) an environment (the landscape the robots navigate), (b) an embodied task (navigation, foraging, construction, etc.) and (c) a decentralized behaviour (each robot acts according to internal computation of local sensing and interactions). Within this framework, language games can, for example, empower the robots to name salient points in the environment (Miletitch et al. 2022), which can later serve to give direction or instructions (Cambier et al. 2023).

Swarm robotics has a strong focus on self-organization, which it shares with evolutionary linguistics (Nolfi and Mirolli 2009). In both cases, agents interact in relatively simple ways to create a complex (physical or cultural) structure. It is therefore relevant to consider a case where interactions between robots in the swarm are language-based (Trianni et al. 2016), if only to study cultural evolution in a dynamic network, that is subject to phenomena such as communication noise (a fundamental feature of communication). On a deeper level, language is fundamentally interwoven with activities (Wittgenstein 1953). As embodied tasks are a core part of swarm robotics, the latter could considerably enrich language evolution models. This is the case for evolutionary

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

approaches as well as cultural evolution models (e.g. language games), as survival, and therefore evolution, depends on successfully achieving environmentally driven tasks, such as foraging, escaping predators, etc. As such, swarm robotics can help answer questions such as:

- How does language emerge on a large scale in a realistic environment?
- To what extent can the emergence of language be driven by use and need?
- How can agents anchor words to a shared environment and agree upon them, using only their individual perceptions, and create meaning through that process?

In this chapter, we present a swarm robotics approach to modelling language evolution. Large-scale robot populations allow for greater opportunities to record data at the level of both the group and the individual. We can study the evolution of the position of the robots in their environment, how efficient they are at tackling their task, specific inner states of the behaviour, as well as information about the status of the evolution of a population-wide language. Due to the complexity and price of handling large groups of physical robots, experiments can also be performed in simulation. In these cases, it is important to be as close to real as possible (physical engine, sensing and communication noise, etc.).

Following this introduction, we start in section 20.2 by detailing the core elements of the swarm robotics methodology, as well as reviewing relevant works in the field. Building on this work, we have developed a swarm robotics model to study language evolution. In section 20.3, we describe necessary steps to make swarm robots fit for the study of language evolution, such as individual relationships and a prosociality factor. These simple adaptations already provide important insights regarding the formation of a collective identity. Notably, this model enables the emergence of in-group bias, i.e. agents are more likely to positively interact with members of their own group. Finally, in section 20.4 we report on experiments combining the setup of the previous section with a collective task, in order to study the impact of the latter on the social and linguistic dynamics. We present our conclusions in section 20.5.

20.2 Swarm robotics: Background

Grasping the core features of the swarm robotics methodology is paramount to designing insightful experiments. We summarize the most important points of the methodology in section 20.2.1. For more details, we refer the reader to more complete reviews of the field, such as Brambilla et al. (2013); Trianni and Campo (2015); Valentini et al. (2017); Majid et al. (2022). With those points laid down, we present, in section 20.2.2, a history of works regarding the evolution of communication and language in swarm robotics.

9780192886491_Book.indb 372 27-Mar-25 23:04:08

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

20.2.1 Methodology

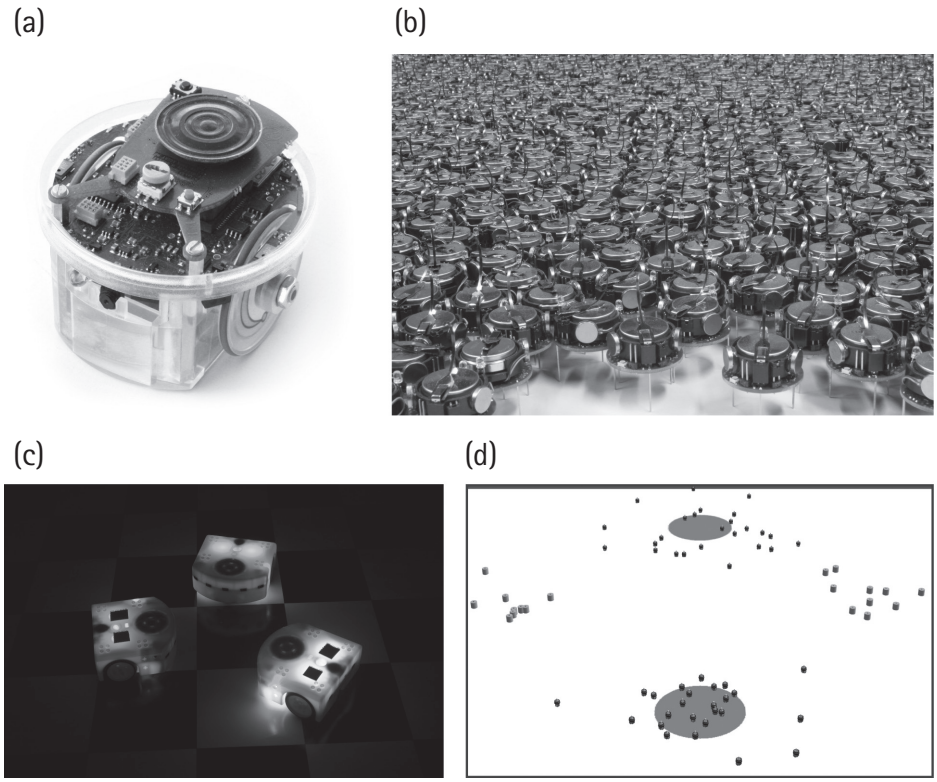


FIGURE 20.1 Swarm robotic experimental platforms.

(a) An e-puck robot.

(b) A swarm of kilobots.

(c) A group of thymios in conversation.

(d) Swarm robotic experiment simulated in ARGoS.

Sources: (a) Wikimedia Commons https://commons.wikimedia.org/wiki/File:E-puck-mobile-robot-photo.jpg. By Stèphane Magnenat. Reproduced under a CC BY-SA 3.0 license; (b) Wikimedia Commons https://commons.wikimedia.org/wiki/File:Kilobot_robot_swarm.JPG. By asuscreative. Reproduced under a CC BY-SA 3.0 license; (c) Computational Intelligence Group, Vrije Universiteit Amsterdam. Reproduced with permission.

In swarm robotics, multiple robots cooperate in order to tackle a given task. The main advantage of using several robots, rather than a single very capable one, is robustness. Indeed, if any single part of the latter breaks, it becomes unable to operate and therefore has to be retrieved and repaired. On the opposite side, with a multi-agent system, such as a robot swarm, if a robot is disabled, the rest of the system should be able to continue operating almost seamlessly. This is especially true in swarms as the agents' behaviour is highly redundant and decisions are taken in a decentralized manner, rather than by a leader (Brambilla et al. 2013).

9780192886491_Book.indb 373 27-Mar-25 23:04:09

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

In many environments, long-range communication is unavailable, thus preventing human control (which would be non-trivial with large quantities of robots anyway) and access to global knowledge (such as global positioning). Swarm robots are therefore autonomous and only rely on local sensing and communication, which improves their scalability and flexibility; the collective behaviour is minimally affected by variations in swarm size or environment, as these changes do not affect the local context.

By virtue of being local, communication is also situated, i.e. the context of the message matters as well as its content (by opposition to abstract communication, whereby only the content of a message is relevant) (Campo 2011). For example, a message containing only a single number can seem useless, however coupled with its context (i.e. the receiver knows this message comes from a nearby robot), it provides information about the local area. In this instance, the number could be a sensor reading indicating the quality of a nearby site.

Robots can therefore use various forms of situated communication to achieve a collective behaviour. For example, they can mark the environment which provides information to their peers when they, in turn, pass through the area (Garnier et al. 2007). This is indirect communication, or stigmergy, and is similar to ants depositing pheromones to find a path in the environment (Deneubourg et al. 1990), or termites strategically depositing mud balls resulting in complex constructs (Grassé 1959). Alternatively, robots might communicate implicitly through, e.g. passive sensing (the quantity of robots detected indicates the popularity of the area; Correll and Martinoli 2011) or pushing and pulling to negotiate the direction of motion (Groß et al. 2006). Finally, direct communication is generally the preferred option as it requires less specific sensors/actuators and can also emulate other forms of communication (Campo et al. 2010; Ferrante et al. 2013). Within this chapter, we focus on this last option, as it is most relevant to modelling language.

These forms of communication enable robots to cooperate in order to achieve a collective behaviour even though individual robots may not perceive it themselves. The behaviours that may emerge in this manner can be categorized into five classes (Trianni and Campo 2015): *aggregation* (i.e. gathering in the same area), *synchronization*, *collective motion* (e.g. flocking or moving whilst maintaining an assembled structure), *collective exploration* (such as the ants example above), or *collective decision-making*. These behaviours can be combined in order to achieve various tasks. For example, in a foraging scenario, robots may explore an environment collectively in order to discover 'food' sources, and use collective decision-making to select the source to exploit first. These behaviours may be performed sequentially or simultaneously. Indeed, approaches that exclusively enable collective decision-making exist (Valentini et al. 2015; Reina et al. 2018), but the latter may also be achieved as a side-effect of other behaviours such as aggregation (Firat et al. 2018) or collective exploration (Gutiérrez et al. 2010).

Achieving these behaviours is, however, not straightforward, as swarm robots are simplistic by design. Indeed, manufacturing and deploying large quantities of robots can only be achieved if the latter are relatively basic. For this reason, swarm robots have limited capabilities regarding sensing, locomotion, computation, and memory.

9780192886491_Book.indb 374 27-Mar-25 23:04:09

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Many swarm robots have been created in the past couple of decades. Noteworthy models include:

- E-pucks (Figure 20.1a): 75*mm*-wide wheeled robots (Mondada et al. 2009). They are equipped with multiple sensors allowing them to detect, among others, the distance and relative position of obstacles, and the colour of the ground. A camera also enables basic colour blob detection, which the robots can use to detect each other as they can light up LEDs on their chassis. Importantly, the e-pucks are compatible with an infrared communication system especially designed for situated communication (Gutiérrez et al. 2008); each message received through this system is associated to a range (i.e. distance with sender) and a bearing (i.e. relative angle between the receiver's heading and the sender's direction).
- Thymios (Figure 20.1c): Educational robots which are easy to program (Riedo et al. 2013). They come with a large array of sensors, although, unlike the e-pucks, none of these sensors are omnidirectional. These limited fields of view can interfere with some standard collective behaviours as many of those assume that local information perception is isotropic (i.e. invariant with respect to direction: all around the robot and not just in front of it). Nevertheless, Thymio can be augmented with additional sensors and computation as it is linkable with an external microprocessor such as a Raspberry Pi.
- Kilobots (Figure 20.1b): inexpensive robots with a diameter of 33 mm (Rubenstein et al. 2012). They are designed to be as cheap as possible so that experiments with potentially thousands of robots are possible. This is necessary to understand collective behaviours at large scales. Despite their extreme simplicity, kilobots have basic communication capabilities.

Setting up experiments with large quantities of robots is time-consuming, and experiments themselves can last for large amounts of time as collective behaviours are highly stochastic. As a consequence, simulators have been developed, which can reproduce the dynamics of swarm robots while streamlining the design and experimental process. A notable example of such a simulator is ARGoS[1] (Pinciroli et al. 2012), which includes models of most swarm robots (including those mentioned above). Other examples include Gazebo,[2] Webots,[3] and CoppeliaSim,[4] which are generic robotic simulators. As embodiment is such an important feature of swarm robotics, such simulators are built around a physics engine that can emulate realistic physical interactions and sensor noise.

The limitations imposed on swarm robots can also be beneficial from a design perspective as multi-agent interactions are complex to model, and their outcomes are

[1] https://argos-sim.info.
[2] https://staging.gazebosim.org.
[3] https://cyberbotics.com.
[4] https://www.coppeliarobotics.com.

9780192886491_Book.indb 375 27-Mar-25 23:04:09

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

therefore hard to predict (Reina et al. 2015). Indeed, differently from a single robot, whose actions can be programmed sequentially to achieve a task with certainty, swarm robots have to be programmed so that their interactions, with each other and with the environment, create an emerging behaviour that achieves the task, but is not, itself, programmable. In other words, there is a gap between the micro-level (i.e. the controller of the robots, which can be programmed), and the macro-level (i.e. the task, which can be defined with precision), that designers can only affect indirectly. From this perspective, the simpler the robots and their interactions, are, the easier collective behaviours are to design and to adjust to a specific task.

Design approaches generally address this problem in either of two ways (Brambilla et al. 2013). On one hand, automatic designs start from the macro-level and devise a numerical definition of the success in addressing the task, i.e. a fitness function. Then, they optimize the robots' controllers (typically, but not exclusively, a neural network), through a large amount of simulated runs, using some meta-algorithm, such as genetic evolution (Trianni 2008). On the other hand, manual approaches are designed from the micro-level as designers start from individual behaviours, which they have reasons to think might achieve the task at hand, often as observed in nature. Then, they iteratively adjust their approach until it reliably addresses the task. This approach often relies on Probabilistic Finite State Machine (PFSM), as exemplified in Figure 20.4 in section 20.4, i.e. a set of individual behaviours which the robot might change to and from with some probability and/or according to local conditions. Ultimately, manual designs have yielded a large range of bio-inspired models of collective behaviours, which can also be useful for artificial life studies, such as those regarding the evolution of language.

20.2.2 Related works

Early works concerned with the emergence of communication within robotic swarms focused on automatic designs. Controllers are evolved, on a specific task, for robots equipped with communication devices (Floreano et al. 2007; Ampatzis et al. 2008). These works focus on questions regarding the conditions that enable the emergence of communication, as communicating is not explicitly rewarded.

In Floreano et al. (2007), the authors evolved two populations of robots. The first was deprived of any communication abilities, while the second was equipped with LEDs and omnidirectional cameras. The robots were placed in an environment with two food sources, one edible and one poisonous, and no means to distinguish one from the other. The robots were evolved to maximize their food consumption. Ultimately, swarms that were able to signal each other performed better. Importantly, the authors assessed various reproduction schemes in the evolutionary process by varying the selection of reproducing agents and relatedness of the offspring. They found that populations selected at the individual level rather than the colony level evolved a disadvantageous communication system. The authors conclude that populations selected

9780192886491_Book.indb 376 27-Mar-25 23:04:09

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

at the colony level created sufficient conditions for the emergence of a cooperative communication.

Multiple studies focused their efforts on engineering, rather than artificial life, asking the question: can robots evolve communication that makes them more efficient? In addition to basic signalling schemes, which can enable or improve collective behaviours such as synchronization (Trianni and Dorigo 2006), coordination (Tuci 2009), or aggregation (Hasselmann et al. 2018), communication can also enable robots to categorize environments in order to adapt the behaviour of the swarm to current conditions (Ampatzis et al. 2008). These works show that communication systems can emerge spontaneously, even if it is not explicitly rewarded by the evolutionary process, as it is beneficial to the task. However, in most cases, the emerged communications are extremely basic and correspond to a simplistic view of signalling, whereby a fixed signal triggers a hardwired reaction in neighbouring agents (Owren et al. 2010). As a matter of fact, in most of the aforementioned works, the signal produced by an individual robot altered the latter's own behaviour as well as its neighbours'. Moreover, these approaches do not take advantage of the self-organization capabilities of swarms, and rather introduce a large overhead time, which prevents the communication system from adapting to new conditions.

On the contrary, in cultural approaches, robots can evolve over the course of the experiment through interactions with one another. Because of their limited capabilities (as mentioned in section 20.2.1), swarm robotics studies on cultural evolution favoured the Minimal Naming Game (MNG) (Baronchelli 2016), a version of the Naming Game that requires a minimal amount of memory and can function without directed communication, i.e. robots do not need to know who they are talking to. Figure 20.2 details the functioning of this model.

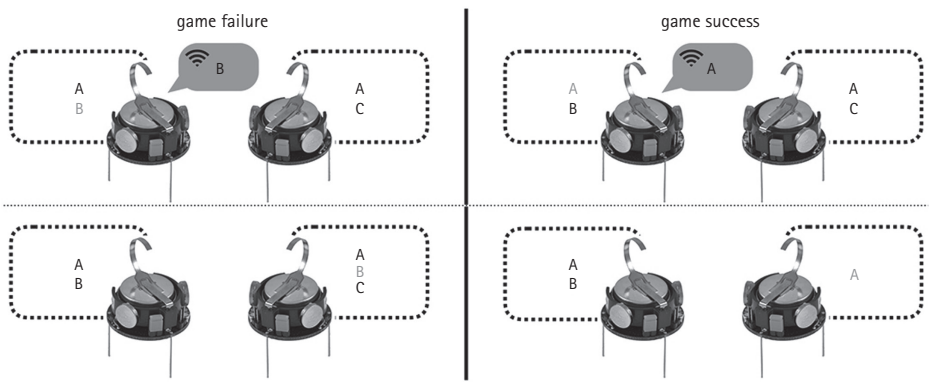


FIGURE 20.2 Example of two Kilobot robots playing the MNG.

The speaker (left robot) chooses one of the words from its inventory and broadcasts it to all the other robots in range. The hearer (right robot) checks the existence of the word in its inventory and updates the latter. Two outcomes are possible, according to whether the robot already knows the word. If the word is absent from the lexicon (left), the word is included into the inventory. If the word already is in the lexicon (right), the word is maintained in the inventory and all other words are removed.

Source: Cambier et al. (2020). Reproduced under a CC-BY license.

9780192886491_Book.indb 377 27-Mar-25 23:04:10

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Once embedded into a swarm performing random motion, the MNG requires even less memory, as communication noise corrupts a significant proportion of sent messages, which reduces the rate of games played and, consequently, reduces convergence time (Trianni et al. 2016). Moreover, the same study showed that variations on the robots' motion (i.e. frequency of turns, which impacts the diffusion of the swarm) could have more impact on convergence dynamics than the rate of speaking in some conditions. This is a clear illustration of the importance of embodiment for cultural evolution. Subsequent works aimed to combine the MNG with more structured collective behaviours. In Cambier et al. (2017), robots are playing an MNG and aggregating only with robots with whom they win at the game. In this case, the collective behaviour and the MNG influence each other, as, conversely, aggregated robots form independent networks that converge to different words. This feedback loop reinforces local aggregates and provides higher control on the collective behaviour as it enables the user to modulate the eventual size of aggregates. Furthermore, as each aggregate converges to a different word, each of those words could later be used as an identifier for the subswarm thus created. Further works on aggregation showed that cultural dynamics could select optimal parameters for aggregation, over the course of a single experiment, without it being explicitly rewarded (Cambier et al. 2021).

The MNG was also combined with collective exploration over a foraging task (Miletitch et al. 2022). Within this process, robots were able to create an accurate description of the environment by collectively naming food sources, and maintaining relevant names over the course of the experiment. A similar process was used to name a large quantity of salient features in an unstructured environment (Cambier et al. 2023). Differently from previous works the names thus created were directly used for the foraging task, as they enabled robots to update each other about the area they visited, and to find the shortest path between their nest and a food source. In addition to benefitting robotics by providing a decentralized path-searching algorithm that requires no prior information about the environment, this study brings forth new insights about the relationship between space conceptualization and language. Indeed, robots are able to find the shortest path between two points without individually mapping their environment.

20.3 Bridging the gap

Current swarm robotics models lack several essential features necessary for this process. Typically, swarm robots are collaborative but homogeneous and with limited memory, and in particular no social memory (such as tracking individual interactions). In this section, we discuss the methods of previous research (Cambier et al. 2022) that aimed to develop an embodied model to study the impact of prosociality on language development using swarm robots.

9780192886491_Book.indb 378 27-Mar-25 23:04:10

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

20.3.1 Prosociality

Previous works assumed the MNG to converge to a uniform state for a well-mixed swarm of robots (Miletitch et al. 2022), as robot swarms tend to have homogeneous behaviours and interactions (see section 20.2.1). However, as the present research moves from an insect-like population to a primate-like one, homogeneity cannot hold as an assumption anymore. To exhibit primate-like mechanics, two features are required: robot individuation, where robots have partner-specific memory (keeping track of the outcomes of past interactions with specific robots), and prosociality, where robots' tendency to interact depends on past experience (successful communication reduces aggression and increases the likelihood of future interactions).

In this model, prosociality is defined as a probability s_{ij} to interact between two distinct robots (i and j). Any robot i calculates s_{ij} based on past interactions with j and a specific equation:

$$s_{ij}' = k_{ij} * \frac{CG_{win}^{ij}}{CG_{win}^{ij} + CG_{lost}^{ij}} + (1 - k_{ij}) * s_{ij} \quad (1)$$

where $CG_{win/lost}^{ij}$ is the quantity of games played between i and j that were won/lost. An initial value s for s_{ij} is required, which is our initial prosociality parameter. k_{ij} is the weighing factor and is given by equation:

$$k_{ij} = min\left\{1, \frac{CG_{win}^{ij} + CG_{lost}^{ij}}{CG_{ref}}\right\} \quad (2)$$

with $CG_{ref} = 10$, allowing for a slow transition from the initial prosociality to the acquired one.

When receiving a message from robot s_j, a robot s_i accepts the interaction with a probability equal to s_{ij}: it hears the word, plays the MNG normally, and updates s_{ij} accordingly. Otherwise, it rejects the interaction and *attacks* robot j to mark its territory, causing robot j to return to its nest.

20.3.2 Experimental setup

This model is tested in simulation, using fifty robots in a 10 x 10 m arena spread equally among two nests. Robots engage in an exploration task (random walk) and play a naming game influenced by prosociality (for a complete description, please see Cambier et al. 2022).

9780192886491_Book.indb 379 27-Mar-25 23:04:14

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

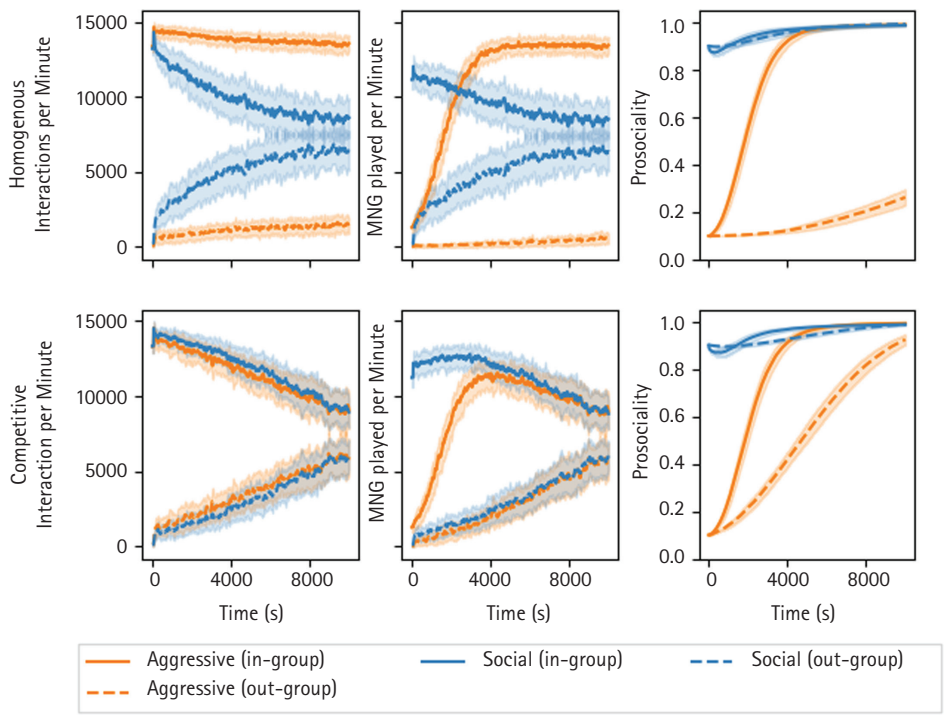


FIGURE 20.3 From left to right, the figure shows the evolution over time of the number of interactions between robots, the number of MNGs conducted, and the value of in- and out-group prosociality.

The top row represents the homogeneous scenario, the bottom row the competitive one. Notably, in the homogeneous context, data for aggressive and social robots come from separate experiments, while in the competitive context, they are part of the same experiments.

Source: Reproduced from Cambier et al. (2024), under a CC-BY 4.0 License.

Most research in swarm robotics focuses on homogeneous swarms, which does not reflect the complexity and variability of the world, which is why we have here two pro-social scenarios:

- a **homogeneous** scenario, where all robots have the same initial prosociality (either low for two aggressive nests or high for two social nests);
- a **competitive** scenario, where each nest starts with different prosociality values (one nest is social and the other is aggressive).

We define in-group prosociality as the mean prosociality value associated with robots within a nest, and out-group prosociality as the mean prosociality value associated with robots from the other nest.

20.3.3 Results

Figure 20.3 displays the number of interactions, and of MNG, and the development of prosociality in both homogeneous and competitive settings.

In the **homogeneous** scenario (top row of Figure 20.3), each robot (either cooperative or aggressive) begins with a high number of in-group interactions and a low number of

9780192886491_Book.indb 380 27-Mar-25 23:04:14

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

out-group interactions. Aggressive robots (i.e. low initial prosociality) tend to provoke rejections (unsuccessful interactions) more frequently than cooperative ones (i.e. high initial prosociality), causing them to return to their nests more often. Conversely, cooperative robots do not retreat to their nests as much and have more chances to interact with other robots, leading to a decrease in in-group interactions and an increase in out-group interactions over time.

Cooperative robots have a higher likelihood of engaging in MNG during interactions, resulting in similar patterns for both the number of interactions and the number of MNG. In contrast, the low prosociality of aggressive robots leads to the rejection of most initial interactions, regardless of the potential MNG outcome. As their in-group prosociality grows over time, more MNG are conducted.

The evolution of prosociality reveals that cooperative robots initially become more aggressive with their kin since they have not culturally converged on a shared vocabulary. Soon, they begin cooperating among each other and eventually with aggressive robots. On the other hand, the latter establish a distinct in-group bias due to their limited interactions outside their nest.

In the **competitive** scenario (bottom row of Figure 20.3), the overall number of interactions decreases over time for both social and aggressive groups, similar to the homogeneous situation for social robots. This can be attributed to the fact that social robots generally initiate more interactions on average, exerting a more pronounced influence on the social dynamics over aggressive robots. We observe similar declining trends for MNG, along with an initial increase in prosociality for the aggressive group, bolstered by interactions with social robots.

The evolution of in- and out-group prosociality, based on swarms' initial prosociality, exhibits a pattern resembling that observed in the homogeneous context. The key difference is that the out-group prosociality of aggressive robots increases more rapidly. This can be explained by the fact that in the competitive scenario, the out-group of aggressive robots is, in fact, composed of social robots—more inclined to engage with the aggressive robots.

20.4 Task-driven language evolution

The model presented in section 20.3 enables the exploration of language evolution theories. This section expands the aforementioned model in order to investigate the human self-domestication hypothesis (HSD), which states that the evolution of present-day languages might have resulted in part from the self-domestication of the human species (Thomas and Kirby 2018; Bentez-Berraco and Progovac 2020). This evolutionary process, which is similar to animal domestication (Hare 2017), would result in less aggressive individuals, more prone to interact with one another (and particularly, with their kin), promoting more social contacts within a community, and supporting the emergence of more sophisticated forms of language (Tamariz and Kirby 2016; Steels 2017).

Self-domestication emerged in a context where cooperation and competitiveness were closely tied to survival and specific tasks, such as foraging. Foraging and resource collection are common animal activities that have been widely studied in swarm robotics (e.g. Miletitch et al. 2018), making them important indicators of evolutionary fitness relevant to the HSD investigation.

9780192886491_Book.indb 381 27-Mar-25 23:04:14

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Self-domestication processes have been associated with two distinct environments: abundant environments with rich and reliable food supplies that can reduce survival pressures, favouring prosocial behaviours over aggression and competition (Hare et al. 2012; Pisor and Surbeck 2019); and harsh environments with limited resources, which can also promote tolerance and friendly behaviour as individuals must rely on non-local food sources to manage resource shortfall risks (Spikins et al. 2021).

Building on top of the experimental setup described in section 20.3, this research aims to model a foraging task in setups recalling the abundant/harsh environments. In this model, robots are tackling a foraging task, played in parallel to prosociality influenced MNG. The goal here is to study the impact of resource distribution and prosocial patterns on the efficiency of foraging. The model and results presented in this section are drawn from Cambier et al. (2024).

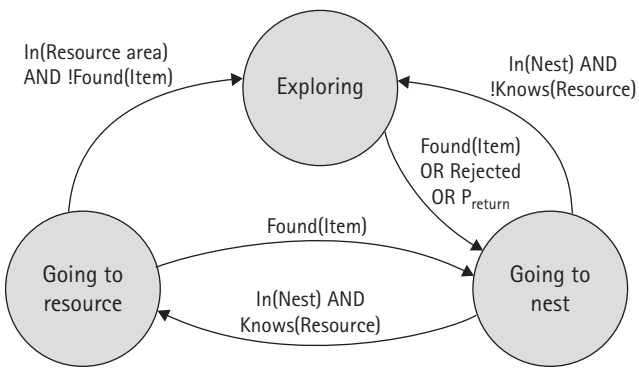


FIGURE 20.4 Probabilistic Finite State Machine guiding the foraging behaviour of our robots.

20.4.1 Foraging

We implemented a PFSM, shown in Figure 20.4, which guides the robots' behaviour. This PFSM is composed of three states:

- **Exploring** the robot conducts a random walk to explore its surroundings. If it discovers an item, it retrieves it and proceeds to the nest. Alternatively, the robot returns to the nest with a probability of $P_{return} = 0.0001$.
- **Going to nest** the robot navigates towards its nest. Upon arrival, it either returns to a known resource location stored in its memory or resumes exploration if no resource location is known.
- **Going to resource** the robot moves towards a known resource position. If it reaches the location without finding an item, it reverts to exploration. However, if it encounters an item, it retrieves it and transports it back to the nest.

Notably, this PFSM requires the robot to remember landmark positions and to navigate towards them (for instance through social odometry—Gutiérrez et al. 2009; Miletitch et al. 2013—a method where robots share and combine their positional data with each other to achieve more accurate self-localization and navigation).

9780192886491_Book.indb 382 27-Mar-25 23:04:15

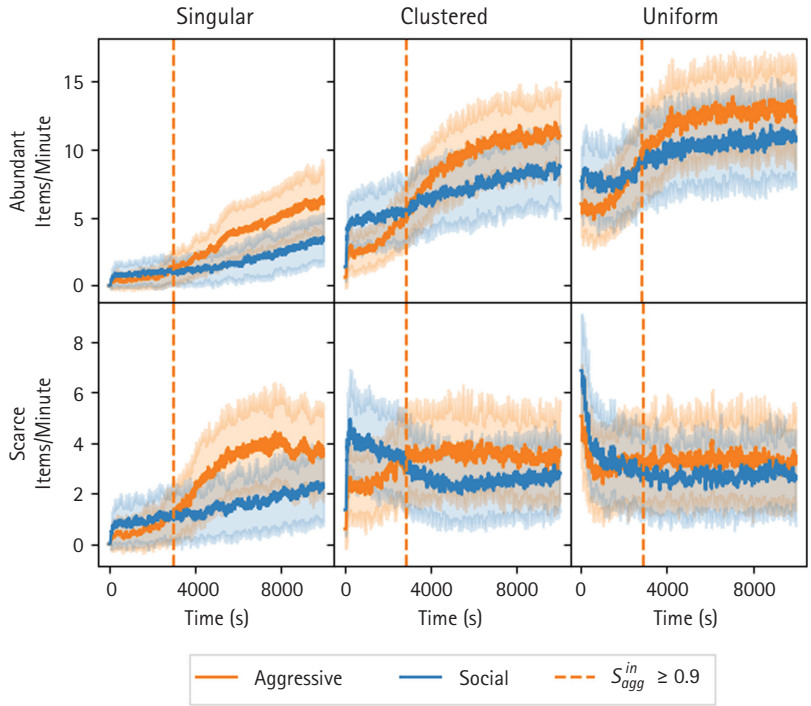


FIGURE 20.5 Evolution of the rate of collected resource items over time (average per nest) for different item distributions (columns), item availability (rows) and initial prosocialities (colours). The nests have the same initial prosocialities.

The dashed orange line indicates the time at which aggressive agents (initial prosociality = 0.1) reach an in-group prosociality of 0.9.

Source: Reproduced from Cambier et al. (2024), under a CC-BY 4.0 License.

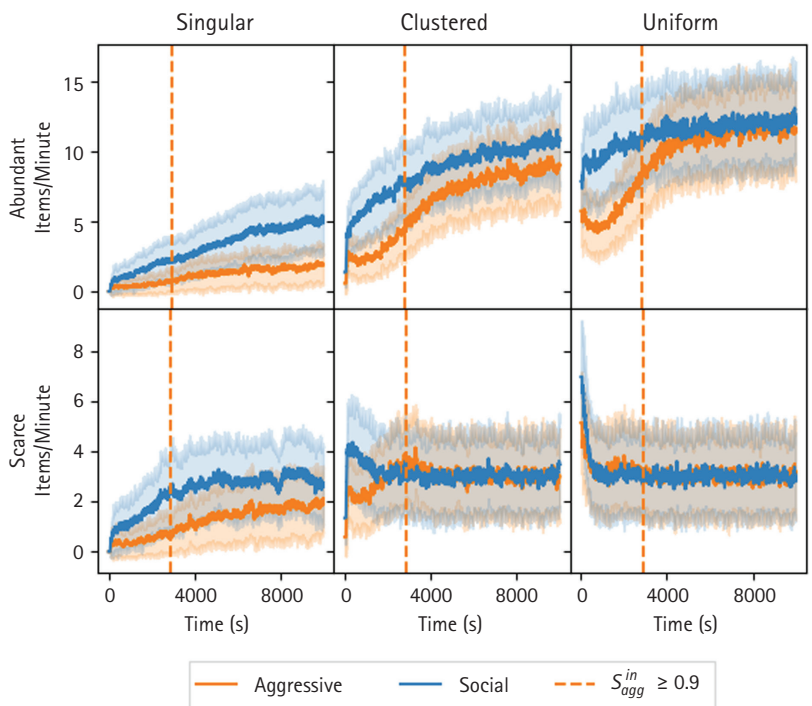


FIGURE 20.6 Evolution of the rate of collected resource items over time (average per nest) for different item distributions (columns), item availability (rows), and initial prosocialities (colours).

The nests have different initial prosocialities. The dashed orange line indicates the time at which aggressive agents (initial prosociality = 0.1) reach an in-group prosociality of 0.9.

Source: Reproduced from Cambier et al. (2024), under a CC-BY 4.0 License.

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

As illustrated in Figure 20.1d, experiments were run with fifty robots, enclosed in a 10 × 10 m arena with two nests, and two resources, from which items spawned with different spreads. Resources are equally distant from each nest, and 2 metres apart from one another. The item distribution of each resource follows a Gaussian curve centred on the resource, with a standard deviation $\sigma \in \{0.01, 1, 10\}$, respectively named *singular*, *clustered*, and *uniform*. Each resource can contain a maximum of twenty-five items, and any time it's not full, items will spawn with a probability $p_{item} \in \{1, 0.005\}$, respectively named *abundant* and *scarce*.

20.4.2 Results

In the homogeneous scenario (Figure 20.5), the average rate of item collection per nest varies based on resource distribution, as the latter leads to different group topologies. In abundant environments (first row of Figure 20.5), where resource exploitation is not limited, and robots may thus continuously exploit the same resources and use the same path, more interactions occur. As aggressive robots use interactions to inhibit exploration, they impede each other and are less efficient than cooperative robots. In scarce environments with clustered or uniform distributions, the rate of resource exploitation is low enough to negate that limitation from aggressiveness.

In the competitive scenario shown in Figure 20.6, social swarms initially perform better at resource collection in abundant environments due to their high initial sociality, facilitating early exploration. However, this advantage is limited to situations where items are easily discovered. Aggressive agents follow a transition function, where a pivot point corresponds to in-group cooperation being achieved. Before this point, aggressive agents underperform as failed interactions force them to return to the nest. Once vocabulary convergence occurs, aggressive robots become equally effective at exploration and collection as social agents but with the added advantage of enforcing a foraging territory by rejecting social robots. This dynamic is especially pronounced in the scarce-singular case, where resource scarcity and ease of guarding resources further favour aggressive agents, hindering social robots' foraging efficiency. The impact of this enforcement diminishes over time as aggressive robots' out-group prosociality increases.

20.4.3 Discussion

In abundant environments, characterized by rich food supplies and minimal competition, social agents demonstrated a higher level of efficiency compared to their aggressive counterparts, in cases where they did not share the same environment. This increased efficiency can be attributed to their ability to exploit resources without the constraints of territorial impediments, which are typically enforced by aggressive agents. The absence of such territorial boundaries allows for better utilization of available resources

9780192886491_Book.indb 384 27-Mar-25 23:04:17

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

and promotes cooperation among the social agents, leading to improved overall performance in foraging tasks.

Conversely, in harsh environments where resources are scarce, or in environments shared with aggressive agents, the advantage of prosociality is limited in the short term. In these challenging conditions, the initial benefits of cooperation and tolerance among social agents are outweighed by the necessity to compete for limited resources. While social agents may initially outperform aggressive agents, their behaviour becomes less advantageous as the competition for resources intensifies over time.

These findings align with the anticipated outcomes of self-domestication concerning behaviour in the two distinct environments. In abundant environments, social agents excel by exploiting resources without territorial constraints. In harsh environments, despite an initial short-term disadvantage, prosociality also provides long-term benefits.

20.5 Conclusion

In this chapter, we have presented an innovative approach to modelling the origin of linguistic systems within the context of swarm robotics. Our framework integrates cultural aspects through a task-driven language game, establishing a foundation for exploring the interplay between self-domestication, language evolution, and swarm behaviour. This is a continuation of early works on language evolution in embodied agents (Steels 2003), that allows studies on very large populations.

Our approach opens up new avenues for research in both theoretical and applied domains. By incorporating individual prosociality and relationships into swarm robotics models, we underscore the potential significance of intricate social networks and the strength of individual connections in shaping collective behaviour. Examining social and linguistic phenomena through the lens of swarm robotics has the potential to garner valuable insights into the emergence of prosocial behaviour, language development, and the evolution of collective intelligence. Our framework serves as a versatile tool that can be adapted and extended to explore a wide range of research questions, making it a valuable resource for other researchers interested in investigating similar social and linguistic phenomena.

One of the key interests of our approach is its potential for the development of self-organizing languages and cultures in swarm robotic systems. Such structure can enable robots to define and execute sequences of tasks through local knowledge sharing, without relying on pre-defined structures or plans. This adaptability is crucial for addressing complex tasks that require multiple actions to be scheduled and coordinated among robots. Our methodology can serve as a foundation for future research exploring the development of fluid construction grammars (Steels 2019) and the advancement of language games in swarm robotics systems.

Additionally, our approach serves as a tool to learn more about natural systems, providing insights into how language, social interactions, and culture function in the

9780192886491_Book.indb 385 27-Mar-25 23:04:17

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

real world for both animals and humans. As such, it not only advances our understanding of robotic communication and coordination but also offers a unique perspective on the study of ourselves and the complex social structures that define our existence.

While our approach adds an extra layer of intricacy to swarm systems because of the per-robot memory needs, we think that this added detail is essential for situations involving complex environments and for the robots to tackle multi-levelled tasks. As models continue to adapt and develop, it will be critical to concurrently create methods and tools for analysing swarm behaviours influenced by an evolving language. This will ensure that our understanding of swarm dynamics keeps pace with the increasing sophistication of swarm robotics systems.

References

Ampatzis, C., Tuci, E., Trianni, V., and Dorigo, M. (2008). 'Evolution of signaling in a multi-robot system: Categorization and communication', *Adaptive Behavior* 16(1): 5–26.

Baronchelli, A. (2016). 'A gentle introduction to the minimal naming game', *Belgian Journal of Linguistics* 30(1): 171–192.

Bentez-Burraco, A. and Progovac, L. (2020). 'A four-stage model for language evolution under the effects of human self-domestication', *Language & Communication* 73: 1–17.

Brambilla, M., Ferrante, E., Birattari, M., and Dorigo, M. (2013). 'Swarm robotics: A review from the swarm engineering perspective', *Swarm Intelligence* 7(1): 1–41.

Cambier, N., Albani, D., Frémont, V., et al. (2021). 'Cultural evolution of probabilistic aggregation in synthetic swarms', *Applied Soft Computing* 113: 108010.

Cambier, N., Eiben, A. E., and Ferrante, E. (2023). 'Emergent naming system in an unstructured environment: A shortest-path discovery case study', in *Artificial Life Conference Proceedings 35*, Vol. 2023, No. 1, p. 47. Cambridge, MA.

Cambier, N., Frémont, V., and Ferrante, E. (2017). 'Group-size regulation in self-organised aggregation through the naming game', in *International Symposium on Swarm Behavior and Bio-Inspired Robotics (SWARM 2017)*, Kyoto.

Cambier, N., Miletitch, R., Frémont, V., et al. (2020). 'Language evolution in swarm robotics: A perspective', *Frontiers in Robotics and AI* 7: 12.

Cambier, N., Miletitch, R., Bentez-Burraco, A., and Raviv, L. (2022). 'Prosociality in swarm robotics: A model to study self-domestication and language evolution', in *The evolution of language: Proceedings of the Joint Conference on Language Evolution (JCoLE)*, 98–100. Nijmegen.

Cambier, N., Miletitch, R., Bentez-Burraco, A., and Raviv, L. (2024). *Pros and Cons of Prosociality across Ecological Niches: A Swarm Robotics Approach*, https://europepmc.org/article/ppr/ppr788240.

Campo, A. (2011). *On the Design of Self-Organized Decision Making in Robot Swarms*. PhD thesis, Université libre de Bruxelles.

Campo, A., Gutiérrez, Á., Nouyan, S., et al. (2010). 'Artificial pheromone for path selection by a foraging swarm of robots', *Biological Cybernetics* 103(5): 339–352.

Correll, N. and Martinoli, A. (2011). 'Modeling and designing self-organized aggregation in a swarm of miniature robots', *The International Journal of Robotics Research* 30(5): 615–626.

Deneubourg, J.-L., Aron, S., Goss, S., and Pasteels, J. M. (1990). 'The self-organizing exploratory pattern of the argentine ant', *Journal of Insect Behavior* 3(2): 159–168.

9780192886491_Book.indb 386 27-Mar-25 23:04:17

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Ferrante, E., Brambilla, M., Birattari, M., and Dorigo, M. (2013). 'Socially-mediated negotiation for obstacle avoidance in collective transport', in *Distributed Autonomous Robotic Systems*, 571–583. Berlin and Heidelberg: Springer.

Firat, Z., Ferrante, E., Cambier, N., and Tuci, E. (2018). 'Self-organised aggregation in swarms of robots with informed robots', in *Theory and Practice of Natural Computing: 7th International Conference, TPNC 2018, Dublin, Ireland, December 12–14, 2018, Proceedings 7*, 49–60. Berlin, Heidelberg: Springer.

Floreano, D., Mitri, S., Magnenat, S., and Keller, L. (2007). 'Evolutionary conditions for the emergence of communication in robots', *Current Biology* 17(6): 514–519.

Garnier, S., Tache, F., Combe, M., et al. (2007). 'Alice in pheromone land: An experimental setup for the study of ant-like robots', in *2007 IEEE Swarm Intelligence Symposium* 37–44. IEEE.

Grassé, P.-P. (1959). 'La reconstruction du nid et les coordinations interindividuelles chezbellicositermes natalensis etcubitermes sp. la théorie de la stigmergie: Essai d'interprétation du comportement des termites constructeurs', *Insectes Sociaux* 6(1): 41–80.

Groß, R., Tuci, E., Dorigo, M., et al. (2006). 'Object transport by modular robots that self-assemble', in *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006*, 2558–2564. IEEE.

Gutiérrez, A., Campo, A., Dorigo, A., et al. (2008). 'An open localization and local communication embodied sensor', *Sensors* 8(11): 7545–7563.

Gutiérrez, Á., Campo, A., Santos, F. C., et al. (2009). 'Social odometry: Imitation based odometry in collective robotics', *International Journal of Advanced Robotic Systems* 6(2): 11.

Gutiérrez, A., Campo, A., Monasterio-Huelin, F., et al. (2010). 'Collective decision-making based on social odometry', *Neural Computing and Applications* 19: 807–823.

Hare, B. (2017). 'Survival of the friendliest: Homo sapiens evolved via selection for prosociality', *Annual Review of Psychology* 68: 155–186.

Hare, B., Wobber, V., and Wrangham, R. (2012). 'The self-domestication hypothesis: Evolution of bonobo psychology is due to selection against aggression', *Animal Behaviour* 83(3): 573–585.

Hasselmann, K., Robert, F., and Birattari, M. (2018). 'Automatic design of communication-based behaviors for robot swarms', in *International Conference on Swarm Intelligence* 16–29. Berlin and Heidelberg: Springer.

Majid, M. H. A., Arshad, M. R., and Mokhtar, R. M. (2022). 'Swarm robotics behaviors and tasks: A technical review', *Control Engineering in Robotics and Industrial Automation: Malaysian Society for Automatic Control Engineers (MACE) Technical Series 2018*, 99–167.

Miletitch, R., Trianni, V., Campo, A., and Dorigo, M. (2013). 'Information aggregation mechanisms in social odometry', in *Proceedings of the 20th European Conference on Artificial Life (ECAL 2013)* 102–109. Cambridge, MA: MIT Press.

Miletitch, R., Dorigo, M., and Trianni, V. (2018). 'Balancing exploitation of renewable resources by a robot swarm', *Swarm Intelligence* 12(4): 307–326.

Miletitch, R., Reina, A., Dorigo, M., and Trianni, V. (2022). 'Emergent naming conventions in a foraging robot swarm', *Swarm Intelligence* 16(3): 211–232.

Mondada, F., Bonani, M., Raemy, X., et al. (2009). 'The e-puck, a robot designed for education in engineering', in *Proceedings of the 9th Conference on Autonomous Robot Systems and Competitions*, Vol. 1, 59–65. IPCB: Instituto Politécnico de Castelo Branco.

Nolfi, S. and Mirolli, M. (2009). *Evolution of Communication and Language in Embodied Agents*. Berlin, Heidelberg: Springer Science & Business Media.

Owren, M. J., Rendall, D., and Ryan, M. J. (2010). 'Redefining animal signaling: Influence versus information in communication', *Biology & Philosophy* 25(5): 755–780.

9780192886491_Book.indb 387 27-Mar-25 23:04:17

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Pinciroli, C., Trianni, V., O'Grady, R., et al. (2012). 'Argos: A modular, parallel, multi-engine simulator for multi-robot systems', *Swarm Intelligence* 6(4): 271–295.

Pisor, A. C. and Surbeck, M. (2019). 'The evolution of intergroup tolerance in nonhuman primates and humans', *Evolutionary Anthropology: Issues, News, and Reviews* 28(4): 210–223.

Reina, A., Miletitch, R., Dorigo, M., and Trianni, V. (2015). 'A quantitative micro–macro link for collective decisions: The shortest path discovery/selection example', *Swarm Intelligence* 9(2-3): 75–102.

Reina, A., Bose, T., Trianni, V., and Marshall, J. A. R. (2018). 'Effects of spatiality on value-sensitive decisions made by robot swarms', in *Distributed Autonomous Robotic Systems: The 13th International Symposium*, 461–473. Berlin, Heidelberg: Springer Science & Business Media.

Riedo, F., Chevalier, M., Magnenat, S., and Mondada, F. (2013). 'Thymio II, a robot that grows wiser with children', in *2013 IEEE Workshop on Advanced Robotics and its Social Impacts*, 187–193. IEEE.

Rubenstein, M., Ahler, C., and Nagpal, R. (2012). 'Kilobot: A low cost scalable robot system for collective behaviors', in *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, 3293–3298. IEEE.

Spikins, P., French, J. C., John-Wood, S., and Dytham, C. (2021). 'Theoretical and methodological approaches to ecological changes, social behaviour and human intergroup tolerance 300,000 to 30,000 bp', *Journal of Archaeological Method and Theory* 28(1): 53–75.

Steels, L. (2003). 'Evolving grounded communication for robots', *Trends in Cognitive Sciences* 7(7): 308–312.

Steels, L. (2017). 'Human language is a culturally evolving system', *Psychonomic Bulletin & Review* 24(1): 190–193.

Steels, L. (2019). 'Basics of fluid construction grammar', in *Case Studies in Fluid Construction Grammar: The Verb Phrase*, 5–52. Amsterdam: John Benjamins Publishing Company.

Tamariz, M. and Kirby, S. (2016). 'The cultural evolution of language', *Current Opinion in Psychology* 8: 37–43.

Thomas, J. and Kirby, S. (2018). 'Self domestication and the evolution of language', *Biology & Philosophy* 33(1): 1–30.

Trianni, V. (2008). *Evolutionary Swarm Robotics: Evolving Self-Organising Behaviours in Groups of Autonomous Robots*, Vol. 108. Berlin, Heidelberg: Springer Science & Business Media.

Trianni, V. and Campo, A. (2015). 'Fundamental collective behaviors in swarm robotics', in *Springer Handbook of Computational Intelligence*, 1377–1394. Berlin, Heidelberg: Springer.

Trianni, V. and Dorigo, M. (2006). 'Self-organisation and communication in groups of simulated and physical robots', *Biological Cybernetics* 95(3): 213–231.

Trianni, V., De Simone, D., Reina, A., and Baronchelli, A. (2016). 'Emergence of consensus in a multi-robot network: From abstract models to empirical validation', *IEEE Robotics and Automation Letters* 1(1): 348–353.

Tuci, E. (2009). 'An investigation of the evolutionary origin of reciprocal communication using simulated autonomous agents', *Biological Cybernetics* 101(3): 183–199.

Valentini, G., Ferrante, E., and Dorigo, M. (2017). 'The best-of-n problem in robot swarms: Formalization, state of the art, and novel perspectives', *Frontiers in Robotics and AI* 4: 9.

9780192886491_Book.indb 388 27-Mar-25 23:04:17

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Valentini, G., Ferrante, E., Hamann, H., and Dorigo, M. (2015). 'Collective decision with 100 kilobots: Speed versus accuracy in binary discrimination problems', *Autonomous Agents and Multi-Agent Systems* 30: 553–580.

Wittgenstein, L. (1953). 'Philosophical investigations, trans', *GEM Anscombe* 261: 49.

9780192886491_Book.indb 389 27-Mar-25 23:04:17

CHAPTER 21

LINGUISTIC EVOLUTION IN TIME AND SPACE

Addressing the methodological challenges

CHUNDRA CATHCART AND BALTHASAR BICKEL

21.1 Introduction

The study of language evolution is an ambiguous project, concerning itself with the biological evolution of the language faculty as well as with the spatiotemporal dynamics of languages over space and time, i.e. what is sometimes called more specifically 'linguistic evolution' (Bickel et al. 2024). Attempts to bridge these two foci propose that at least some aspects of the evolution of the language faculty were driven by the same extragenomic transmission processes that also drive linguistic evolution (e.g. Kirby 2017). In other words, currently observable language change to some extent recapitulates the evolution of the language faculty in the hominin lineage. The central methodological toolkit for this enterprise are experiments on social learning and micro-evolutionary studies on how linguistic structures emerge in iterated populations of users.

Here we focus on an alternative bridge, with different methodological implications (Bickel 2015; Bickel et al. 2023). We probe linguistic evolution as a specific trait of the human phenotype, an intrinsic property of the language faculty. Language change reveals some of the key mechanisms that underlie this faculty, and it therefore points to what exactly must have come together when the faculty emerged—specific social processes of convergence and divergence and specific neural processes of learning, producing, and comprehending, both conditioned on laws of mathematics (e.g. information theory) and (bio)physics (e.g. articulator movement). Examining these mechanisms through evidence from language change expands the purview of language origins research from the static possibilities (such as compositionality) to the dynamic probabilities of the language faculty (such as the preference for minimizing

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

dependencies during language change, or the preference for accelerated divergence when languages split).

The methodological toolkit needed for this enterprise consists primarily of models that capture language change and diversification processes. Interest in questions of this sort directly connects with a long history in the fields of historical linguistics and linguistic typology. In past decades, the advent of probabilistic modelling of linguistic questions has made it possible to formulate explicit testable hypotheses regarding language change. This chapter focuses on Bayesian computational models designed for this purpose. Models of this sort aim to shed light on the processes that give rise to extant linguistic data, taking into account cognitive biases of learning and processing along with social and demographic phenomena such as language contact, as well as interactions between these conditions on linguistic evolution.

In this chapter, we provide a review of statistical modelling of linguistic evolution, focusing on two powerful families of Bayesian models adopted from computational biology that are suited to complementary aims, explicitly characterizing, on one hand, evolutionary pressures in language change and the stability of features, and on the other, contact and admixture between languages. While some attempts have been made to reconcile these two approaches and integrate them with each other in models of vertical transmission and horizontal transfer, these models are beset by a number of limitations.

We introduce the nascent framework of distributional phylogenetic modelling, a flexible family of phylogenetic modelling that allows for the exploration of the effects of certain evolving variables on the patterns of change in others. We outline some ways in which distributional phylogenetic models can incorporate phylogeographic information in order to test hypotheses regarding the relationship between population history and patterns of change. We present a case study that assesses the role of patterns of geographic dispersal in the evolution of phonological inventories in the Austronesian languages.

We find decisive evidence that longer migrations from the Austronesian homeland coincide with faster rates of linguistic change and decreases in inventory size, counter to predictions derived from the literature on sociolinguistic pressures in language change. We discuss the broader implications of these results, highlighting some pitfalls of relying on geospatial information alone as proxies for language contact and stable multilingualism. We discuss ways in which this framework can be further enriched and expanded to shed light on the co-evolution of linguistic and extralinguistic traits.

21.2 Background

Linguistic typology seeks an understanding of the processes giving rise to distributions of linguistic features, with an eye to associations between linguistic features as well as associations between extralinguistic and linguistic features (Bickel 2007). In recent decades, the availability of large linguistic databases (see Ivani and Bickel 2024 for a

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

review of syntactic databases; ASJP—Wichmann et al. 2018—for a lexical database; and PHOIBLE—Moran and McCloy 2019—or LAPsyd—Maddieson et al. 2013—for phonological databases) has made it possible to conduct data-driven investigations into the distribution of linguistic features, quantifying the effects of both historical contingencies ('event-based' factors) and of general principles of language use, learning, production, and comprehension ('functional' factors), as well as the interaction of these (Bickel 2015). Linking large linguistic databases with non-linguistic metadata pertaining to the languages in them has facilitated investigations into a large number of associations between linguistic and extra linguistic features that bear on evolutionary patterns (Nettle 1999; Atkinson 2011; Everett 2013; Bickel et al. 2015; Everett et al. 2015; Roberts 2018; Blasi et al. 2019; Barbieri et al. 2022).

As in other fields, a challenge to modelling the relationship between linguistic features and other linguistic or extralinguistic features is the presence of non-independence among the data points observed due to phylogenetic relatedness, an issue known as Galton's Problem (Narroll 1961). A group of related languages may display a correlation between two linguistic features because there is some functional benefit to their co-occurrence, or because they have simply inherited both features from a common ancestor, in which the features came to co-occur by chance. Properly addressing this issue is key to understanding the role of different pressures in language change and explaining which constellations of features are due to functional pressures and which are due to historical contingencies, or both.

Galton's Problem has been addressed in a variety of ways. The most common strategy takes relatedness as a nuisance factor in need of control. The classical implementation of this (Dryer 1989) was to restrict samples so as to minimize relatedness, e.g. by sampling only one language per genus, a taxonomic level defined as groups roughly comparable to Indo-European subgroups. This approach effectively enforces phylogenetic independence within a dataset, but results in massive underuse of available data, which can result in reduced statistical power. A more recent alternative is hierarchical (alternatively 'mixed-effects' or 'multi-level') regression models (Jaeger et al. 2011). Models of this sort involve partial pooling, assessing the influence of effects of interest on a response variable while accounting for variation at the group level ('random effects') that should not be conflated with the population-level effect of interest ('fixed effects'), e.g. by capturing family- and area-specific biases that might influence fixed effects in models with complete pooling (Gelman and Hill 2007).

A drawback of hierarchical models in which coefficients vary at the family level is that they treat families as discrete units and relatedness among languages as a binary variable. French, Italian, and Pashto are all Indo-European languages, but French and Italian are demonstrably more closely related to each other than they are to Pashto, as they both descend from Latin. This property of relatedness can be captured by hierarchical models in which the propensity for a feature which a language displays due to its phylogenetic relatedness to other languages is generated by a Gaussian Process (Guzmán Naranjo and Becker 2022); language-level feature biases are distributed according to a multivariate Normal distribution with a mean of zero and a covariance

9780192886491_Book.indb 393 27-Mar-25 23:04:17

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

matrix based on the phylogenetic relatedness of languages in the sample. The same process can be used to quantify the expected similarity among languages with respect to some propensity for a feature due to their geographic proximity.

In short, models in this tradition ensure that all potential nuisance factors are controlled for when investigating the relationship between two variables. The benefit is a control of false positives. Controlling for relatedness in hierarchical models guards against mistaking a pattern for independent evidence when it exists just because languages descend from the same ancestor. At the same time, however, these models risk false negatives. When related languages maintain the same pattern since they split from each other, this could just as well point to some external pressure (functional or event-based) that keeps the pattern (Bickel 2015). Hierarchical models absorb any such signal as part of the group variance (random effects), not as effects of interest at the population (fixed-effect) level. Therefore, we might miss signals, and the probability of this increases with increasing phylogenetic depth in the groups we control for.

An alternative approach that has gained considerable traction over the past twenty years is that of phylogenetic linguistics (or 'phylolinguistics': Greenhill et al. 2021; Bayesian Phylolinguistics), which borrows phylogenetic comparative methods (PCMs) from computational biology to the study of linguistic evolution. For the most part, these approaches explicitly model the transmission and diffusion processes thought to give rise to the diversity we observe, rather than treating them as nuisance factors. Another approach to analysing linguistic diversity in this tradition is the use of models which explicitly represent the role of areal admixture in shaping cross-linguistic distributions of features. We describe these families of models in the following section.

21.3 Explicit models of linguistic evolution

21.3.1 Bayesian modelling

Recent years have seen the rise of Bayesian modelling in a number of subfields of linguistics. Bayesian methodologies allow practitioners to create models characterizing stochastic processes capable of generating observed data, expressed in the form of probability distributions (e.g. Normal, Binomial, etc.). This generative process usually results in a tractable likelihood (on models with intractable likelihoods, see Sisson and Fan 2011) of the data under different values of model parameters (unobserved quantities involved in the generative model), written. Model likelihoods can be used to estimate parameter values. In a maximum likelihood framework, this involves finding a single point estimate for θ which maximizes the model likelihood. Bayesian modelling incorporates prior beliefs regarding parameter values to estimate posterior parameter

9780192886491_Book.indb 394 27-Mar-25 23:04:17

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

distributions, which peak around values with high posterior probabilities, according to Bayes' theorem:

$$P(\theta \mid data) = \frac{P(data \mid \theta)P(\theta)}{\int P(data \mid \theta)P(\theta)d\theta}$$

Parameter estimation allocates credibility to different hypotheses regarding the data generation process, as certain parameter values will be compatible with certain hypotheses but not others.

For some simple models, posterior parameter distributions have an analytic solution. In most cases, however, posterior distributions involve an intractable integral and require some form of estimation. Markov chain Monte Carlo (MCMC) is a popular family of sampling-based algorithms for inference of posterior parameters. Parameters are initialized with random values, which are stochastically updated according to different algorithms designed to steer values towards regions of higher posterior probability (Geman and Geman 1984; Robert et al. 2010) for a large number of iterations, until posterior samples are drawn from the stationary distribution of the chain. For a more detailed introduction to Bayesian inference, see Gelman et al. (2013), Stone (2013), Kruschke (2014), and McElreath (2020).

21.3.2 Phylogenetic modelling

A prominent use of phylogenetic methods in linguistics has involved inferring phylogenies of language families. While the traditional comparative method of historical linguistics is capable of establishing secure subgroups within language families on the basis of shared innovations, it cannot always resolve higher order subgrouping within families as well as questions regarding the absolute chronology of divergences. Bayesian phylogenetic inference provides a means of allocating credibility to different candidate tree topologies and estimating divergence dates while incorporating a degree of uncertainty. In many cases, these methods have produced results that dovetail with received wisdom regarding the dispersal of language families, but in other cases, have produced conflicting results (Bouckaert et al. 2012; Chang et al. 2015; Heggarty et al. 2023).

Phylogenetic comparative methods (PCMs) are a family of methods that model the evolutionary dynamics of various features. In general, these methods rely on an existing phylogeny of the organisms or languages under study, on which dynamics of change are conditioned. Likelihood-based PCMs have a number of applications. A popular set of methods quantify the phylogenetic stability of features (Abouheif 1999; Blomberg et al. 2003; Borges et al. 2019) and compute phylogenetically corrected means, and have enjoyed a degree of use in linguistics (Macklin-Cordes et al. 2021; Carling and Cathcart 2021a).

9780192886491_Book.indb 395 27-Mar-25 23:04:18

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

PCMs can be applied to categorical and continuous data types. Most applications in linguistics have so far focused on categorical data because the relevant features are typically discrete, or modelled as discrete (e.g. verb–object vs. object–verb order, labiodental vs. bilabial stops, or descendants of *sem- vs. *Hoy- as the label for 'one' in Indo-European). A popular model for the evolution of categorical linguistic data assumes that features undergo state changes over a phylogeny according to a continuous-time Markov (CTM) process, a stochastic process under which transitions between different states (or feature values) in a system occur according to transition rates, representing the frequency of different changes. Estimated rate values can be used to address hypotheses regarding the dynamics of change of the features in question. The likelihood of a set of rate values, $P(data \mid rates, tree)$ can be efficiently computed by the pruning algorithm (Felsenstein 1981, 2004), a dynamic programme that exploits the independence structure of directed acyclic graphs (see Figure 21.1 and §8.8 for a full walk-through). Posterior distributions for rates can be estimated via an MCMC algorithm. Following inference of rates, character histories can be simulated over the phylogeny using a technique known as stochastic character mapping (Huelsenbeck et al. 2003; Bollback 2006; Revell 2013). This allows graphical visualizations of how probabilities of states gradually change over time (Widmer et al. 2017; Blasi et al. 2019; Cathcart et al. 2020; Jing et al. 2023).

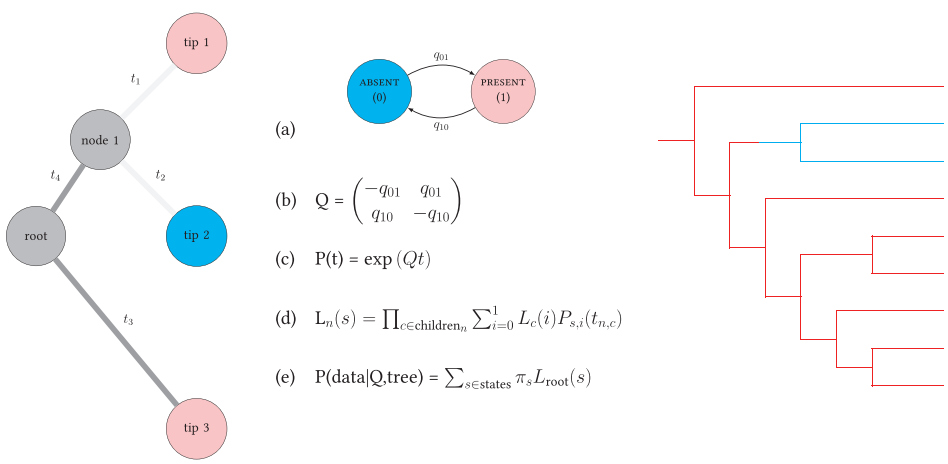


FIGURE 21.1 Illustration of the computation of rate likelihoods under the pruning algorithm (Felsenstein 1981) for a binary character (left-hand figure) and a hypothetical history of a binary character, visualized as a stochastic character map (right-hand figure).

Transitions between states occur according to a gain rate q_{01} and loss rate q_{10} (a). The rate matrix Q (b) can be exponentiated in order to generate a matrix of state transition probabilities $P(t)$ over a given interval of time t (c). For a given node in the tree n, the likelihood of a state s conditioned on the state values of all nodes descending from node n (d) can be computed using these transition probabilities, along with likelihoods at descendant states—for example, the likelihood of node 1 in state 1 (red), $L_1(1)$, is the product of the probability of ending in state 1, $L_{c=\text{tip }1}(1)$, in the t_1-branch, $P_{1,1}(t_1)$, and the probability of ending in state 0, $L_{c=\text{tip }2}(0)$, in the t_2-branch, $P_{1,0}(t_2)$, with $L_c(i)$ normally evaluating to 0 or 1 at the tips (though see Jing et al. 2023 for generalization). These likelihoods are computed recursively in post-order traversal, with parent nodes of lighter branches visited prior to parent nodes of darker branches. The likelihood for the entire tree (e) is equal to the dot product of state likelihoods at the root and π, the vector of prior root state probabilities, for example, drawn from a uniform or the stationary distribution (see Felsenstein 2004; Maddison et al. 2007; FitzJohn et al. 2009; Beaulieu and O'Meara 2014).

9780192886491_Book.indb 396 27-Mar-25 23:04:18

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

The CTM model can be used to address a wide range of questions regarding language change. A CTM chain has a stationary distribution, which represents the proportion of time that the system is expected to be in a particular feature state as time approaches infinity. When the CTM has reached stationarity, this means that there will always be the same proportion of languages in different states even though individual languages keep changing between states. Accordingly, the proportions are independent of any original or current state, i.e. they are in a sense 'timeless'. Assume there are 100 languages in state A and 100 in state B, and stationary transition probabilities of $P(A \rightarrow B) = .2$ and $P(B \rightarrow A) = .8$. A will gain 80 and lose 20; and B will gain 20 and lose 80, resulting in 160 A and 40 B languages. The same result will be achieved when starting from, say, 150 A and 50 B languages (A: $150 - .2 \times 150 + .8 \times 50$; B: $50 - .8 \times 50 + .2 \times 150$), and in fact from any distribution of frequencies. Accordingly, CTM models can be used to estimate long-term preferences for feature states that are independent of time (Maslova 2000; Cysouw 2011; Bickel et al. 2015; Widmer et al. 2017; Jäger and Wahle 2021).

Another use of these models is to investigate evolutionary interdependence between features. A popular model assesses whether changes in the state of one feature depend on values of a second feature (Pagel 1994); this model has enjoyed wide use in linguistics in order to explore the dynamics of change between features that are synchronically highly correlated (Dunn et al. 2011; Cathcart et al. 2020; Jäger and Wahle 2021). Models of this sort have been used to investigate constrained pathways of change (Dunn et al. 2017; Shirtz et al. 2021). Additionally, the CTM model of character evolution is capable of carrying out ancestral state reconstruction to internal nodes of a phylogeny representing unattested protolanguages (Maurits and Griffiths 2014; Zhou and Bowern 2015; Widmer et al. 2017; Blasi et al. 2019; Carling and Cathcart 2021b). While these reconstructions typically fit well with traditional qualitative reconstructions, they have the advantage that they come with uncertainty estimates and can reach much greater time depth than qualitative methods that rely on the documentation of earlier stages (ancient languages).

Despite their productive use in linguistics, CTM models are not without their limitations, at least in their most commonly used implementations. One drawback is that the transitions that can be modelled involve only attested states; for practical purposes, a CTM model can only make inferences regarding transitions between feature values observed at the tips of a phylogeny, and therefore, would not be capable of reconstructing a presently unobserved value. Sometimes this problem can be resolved by recoding features; for example, a word order variable containing values such as SOV, SVO, etc., can be recoded into orthogonal binary features with dynamics that can be inferred independently (SV vs. VO). This provides the opportunities for a wider range of feature combinations to be reconstructed.

Another issue is that fitting CTM models is generally restricted to large language families with some branch length estimates in terms of calibrated calendar years or in terms of amounts of change (substitutions). There have been only a few attempts to generalize the approach to small families and isolates, all seeking to combine information from small and large families in a single model to the maximum extent possible.

9780192886491_Book.indb 397 27-Mar-25 23:04:19

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

One uses stationary probabilities from large families as priors for modelling the probability distributions in isolates and small families synchronically (Bickel et al. 2015; Bickel 2015). Another approach assumes that an isolate is not only an extant language but, at the same time, the root of a degenerate (zero time) phylogeny. Its likelihood is then estimated together with the likelihoods at the root from larger language families (Step (e) in Figure 21.1; Jäger and Wahle 2021). Yet another approach might assume the time depth of isolates to lie between 5,000 years (after which one would detect phylogenetic relations) and a maximum based on major human migration events (before which lineages would stem from the same population). These estimates could then be combined into a global supertree (Bouckaert et al. 2022; Shcherbakova et al. 2023), or, alternatively, into a forest of trees on which a model is fitted.

A further issue is that CTM models work best when targeting a single variable which takes one value per language (and this is how typological databases tend to code languages, apart from a few exceptions; see Ivani and Bickel 2024 for a survey). But languages are complex systems with rampant polymorphism: languages are very often split in terms of word order (e.g. by clause type or information structural condition), case systems (e.g. by tense or aspect, or person), morphological syncretisms (e.g. by which cells they span, or by conjugation class), or even syntactic possibilities (e.g. allowing recursive noun phrases with prepositions but not with adjectivization). There have been various attempts of modelling such splits, tailored to specific data and research questions. Some have treated subsystems as uncertainties, or as if they descended from a language like dialects (Bickel et al. 2015). A more explicit approach models subsystems independently, each on their own, and then captures interactions and dependencies by aggregating posterior probabilities per time interval through stochastic character mapping (Widmer et al. 2017). An approach that models the dependencies between subsystems directly is to include split states into the CTM, so that languages can transition from, say, general nonergative to a split ergative/accusative to a general accusative system, with or without additionally allowing direct transitions (Hong et al. in prep.). Alternatively, one can treat the split condition as a predictor variable for the presence of a subsystem (e.g. different cell collections that each predict the presence of syncretism), and then model the conditional presence in the CTM (cf. Cathcart et al. 2022 and section 21.4).

Splits often come with probability distributions that shift between generations, and this represents yet another challenge for classical CTM approaches. From a microevolutionary perspective, a 'gain' or 'loss' of a linguistic feature represents the point in time when the inflection point of the S-shaped growth curve that characterizes the rise of linguistic variants (Greenberg et al. 1954; Bailey 1973; Niyogi et al. 1997; Yang 2000) is surpassed. What is therefore needed is models that go beyond discrete grammar features and tackle the probabilistic nature of language directly. One recent approach treats variation in word order as the product of probabilistic grammars, and then takes corpus-based probabilities for the likelihood estimation at the tip in the CTM model (Figure 21.1; Jing et al. 2023). Another approach replaces CTM

9780192886491_Book.indb 398 27-Mar-25 23:04:19

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

models with models for continuous measures, such as Brownian Motion or Ornstein-Uhlenbeck models (Witzlack-Makarevich et al. 2016; Hahn and Xu 2022).

Finally, CTM models do not explicitly model contact dynamics; if under a CTM model a feature is likely to have arisen frequently within a phylogeny, a CTM model cannot distinguish which births are due to functional pressures (e.g. adaptation to some processing principle) and event-based triggers where the feature has entered a language due to social conformity. Post-hoc analyses can give some insights, for example by comparing whether the evolutionary dynamics is similar between families in the same contact area vs. those in different areas (Bickel et al. 2015; Bickel 2019). Direct inference of contact effects is rare but there are promising developments in the phylogenetic inference of language trees. Some models are successful in inferring lateral transfer between related languages based on patterns of co-occurrence (Kelly and Nicholls 2017; Neureiter et al. 2022), but these do not (yet) generalize to PCMs and do not take geography or proposed areal hypotheses into account.

21.3.3 Areal modelling

Other research explains variation in cross-linguistic patterns via models that explicitly represent areal influence and admixture between languages. A significant portion of this work makes use of mixture models. Mixture models partition data points in a sample into clusters depending on the values for features exhibited by individual data points, with data points sharing similar values belonging to similar clusters. As such, mixture models provide a means of dimensionality reduction, the resulting representation of which may have an interpretation involving language contact. Some, but not all of this work injects explicit spatial information into the probabilistic model used.

For linguistic purposes, clustering can take place at a number of levels. Languages themselves can be clustered on the basis of the features they display (Michael et al. 2014). Alternatively, the features within languages can be partitioned among different clusters. This is the objective of a subtype of mixture model known as admixture or mixed membership models. Mixed membership models allow for features within a language, words within a document (Blei et al. 2003), or single-nucleotide polymorphisms within genomic data (Pritchard et al. 2000) to be generated by one of a number of latent components. This methodology has been used to model admixture within languages corresponding to different areal sources (Reesink et al. 2009; Bowern 2012; Syrjänen et al. 2016; Cathcart 2020, 2022). Mixed membership models have also been used to ask which of a language's features are due to genetic inheritance as opposed to contact or universal preferences (Chang and Michael 2014), and can thus be used to quantify the borrowability and propensity for inheritance of different features. A recent approach, SBayes (Ranacher et al. 2021), infers the presence of areal groupings across languages. Under the generative process for this model, a language

9780192886491_Book.indb 399 27-Mar-25 23:04:19

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

can be assigned to an areal group according to a spatially informed prior distribution over group membership (the other models cited above do not provide any explicit geographic information to their models). Within a language, each feature is assumed to come from three different sources: inheritance (depending on the distribution of the feature in the family to which the language belongs), contact (depending on the distribution of the feature in the areal group to which the language is assigned), and universal preferences (depending on the global distribution of the feature). These distributions are multinomial probabilities expressing the preference for a given feature value in a given linguistic taxonomy.

While mixture model-based approaches can explicitly model contact in a way that standard phylogenetic methods do not, they lack phylogenetic models' explicit, fine-grained representation of time depth and diachrony. The standard admixture model strongly embodies the Hardy-Weinberg principle in that it assumes that no mutations have taken place since the time of admixture between populations. This assumption may not be valid for all linguistic features. For instance, two languages could share a word order due to a relatively recent change, yet an admixture model could potentially impute this resemblance to deeper contact. One way to allay this concern is to work with features thought to be relatively stable or invariant, such as the operation of sound changes demonstrated to have taken place at an early date (Cathcart 2020, 2022). At least one population genetic model, overlooked in the linguistics literature, allows for a limited number of mutation events representing changes occurring between the time of admixture and the present day (Shringarpure and Xing 2009). Even with this relaxation, these models' representation of chronology is simplistic and not as expressive as that of timed trees. This temporal invariance is also seen in the representation of genetic featural propensities found in SBayes. In SBayes, the distribution of variants of a feature (e.g. possible word orders) in a family such as Indo-European is represented by a single multinomial probability distribution generated by a Dirichlet prior distribution. This flat representation of family-level feature distributions has no sensitivity to the fact that Welsh and Irish are more likely to share a word order given their shared phylogenetic history than Welsh and Pashto. The main benefit of this representation is its contribution to the model's computational tractability. An alternative would be to assume that the log probabilities of different features evolve according to Brownian motion, and that feature probabilities themselves follow the logistic normal distribution, for which there exists a conjugate prior for the multinomial distribution which is however somewhat involved to implement (Mimno et al. 2008).

A number of methods attempt to synthesize phylogenetic and areal models somewhat more directly, although they do not necessarily capture the full evolutionary dynamics of transmission and diffusion. Daumé III (2009) constructs a mixture model designed to tease apart the roles of contact and descent in shaping typological distributions. Under this model, a feature in a language is generated according to either the areal group to which the language is assigned, if it is assigned to one, or via a coalescent process. The author shows that this model's spatial awareness results in the inference of more accurate phylogenies than a model without spatial awareness. At the same time, this model

9780192886491_Book.indb 400 27-Mar-25 23:04:19

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

would be difficult to adapt to questions regarding prehistoric contact, etc. In a similar vein, Murawaki and Yamauchi (2018) employ an autologistic model to infer vertical stability and horizontal transmissibility. While these models advance our understanding of the interplay of different pressures in linguistic evolution, adapting and expanding these models to take into account different sources of variation (and to accommodate different prior distributions) may be a challenge, given their reliance on certain inference algorithms. More critically, information regarding areality tends to be based on present-day distributions, with no readily available way of shedding light on sociolinguistic pressures active at a greater time depth.

21.3.4 Issues regarding implementation

This brings us to yet another point. Coding up inference algorithms for complex probabilistic models from scratch is time-consuming and error-prone, and may be beyond the remit of many linguists. Fortunately, a number of accessible probabilistic programming languages on the market make it possible for analysts to specify the data generating process, while the software does the heavy lifting of posterior parameter inference. A popular language is Stan (Carpenter et al. 2017), which is well documented and has a large online community of users. A limitation of these software packages, which generally make use of gradient-based methods to steer parameter estimates towards regions of high posterior probability, is that they cannot estimate discrete posterior distributions (as they are not differentiable), which must be marginalized out by summing.

This places constraints on the models that can be fitted using such programs, and in some cases, a bit of rethinking such models from a standpoint of tractability. For instance, it would not be possible to directly implement SBayes in Stan. SBayes samples a discrete variable representing the membership for a language with index l in an areal group, which we can call $z_1 \in \{1,\ldots,K\}$, and for each feature with index $d \in \{1,\ldots,D\}$ in the language, samples a discrete variable indicating whether the feature in question is inherited, borrowed, or due to universal preferences, which we can call $s_{1,d} \in \{I,B,U\}$. The likelihood of a configuration for a single language could be written as follows, with simplified notation (θ refers to all model parameters; $P(y_{l,d} \mid s_{l,d})$ is the probability of the observed value of a feature in a language conditioned on its source):

$$P(z_l \mid \theta) \prod_{d=1}^{D} P(s_{l,d} \mid \theta) P(y_{l,d} \mid s_{l,d})$$

In the context of Gibbs sampling with a conjugate prior, the different discrete variables $(z_l; \boldsymbol{s}_{l,\cdot})$ can be sampled conditioned on the variables on which they depend, which include current values for other discrete parameters. If configurations of discrete variables must be marginalized out, we are faced with the following sum:

$$\sum_{k=1}^{K} \sum_{x_1 \in \{I,B,U\}} \ldots \sum_{x_D \in \{I,B,U\}} P(z_l = k)\ P(s_{l,d} = x_1 \mid \theta)\ P(y_{l,d} \mid s_{l,d} = x_1) \ldots P(\boldsymbol{s}_{l,d} = x_D \mid \theta) P(y_{l,d} \mid s_{l,d} = x_D).$$

9780192886491_Book.indb 401 27-Mar-25 23:04:22

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

This sum is intractable as it enumerates $K3^D$ different configurations of values for discrete parameters. The issue is that because $\boldsymbol{s}_{l,.}$ depends on z_l, all possible combinations of this collection of discrete variables must be taken into account. This issue would be alleviated if individual features in a language could belong to different linguistic areas $z_{l,d} : d \in \{1, \ldots, D\}$, altering the above likelihood as follows:

$$\prod_{d=1}^{D} P(z_{l,d} \mid \theta) P(s_{l,d} \mid \theta) P(y_{l,d} \mid s_{l,d})$$

Each term in this product represents the likelihood of a feature in a language, $y_{1,d}$. Summing out the discrete parameters for all of these likelihoods in a single language requires only $3DK$ operations. Ultimately, whether or not they are justified, sacrifices like this one open the door to greater ease of model fitting with more flexibility, including use of diverse, not necessarily conjugate priors.

It is likely that in the near future, the field will see an explosion of models designed in this spirit. We believe that progress in this direction will be facilitated by easily adaptable and expandable models cast in probabilistic programming languages that serve as *lingue franche* for analysts in the field. In the following sections, we outline a program for flexible phylogenetic models implemented in RStan, inspired by advances in hierarchical Bayesian modelling made in the brms framework (Bürkner 2017).

21.4 Distributional phylogenetic modelling

In this section, we present a flexible method for analysing multifaceted aspects of linguistic evolution in a nuanced manner. We term this program distributional phylogenetic modelling. We take as our point of departure two unrelated families of statistical models. The first group consists of phylogenetic models and phylogenetic comparative methods which model rate variation in the evolutionary histories of traits. The second group consists of distributional regression models, an increasingly popular Bayesian modelling framework which allows both the location and scale of a regression model to vary as a function of predictor variables (Bürkner 2017), thus relaxing a number of assumptions found in classical regression, such as homoskedasticity. We demonstrate a novel use of this modelling strategy, highlighting its ability to integrate spatial and temporal aspects of linguistic evolution and shed light on poorly understood aspects of change.

PCMs for discrete features generally assume that a feature moves between different values or states according to a continuous-time Markov process parameterized by transition rates (Figure 21.1). Under the standard view of a continuous-time Markov process for binary data, a feature arises and is lost according to a gain rate and a loss rate. Assuming a speed of change s (loss *or* gain) and stationary probability p, the gain

9780192886491_Book.indb 402 27-Mar-25 23:04:25

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

rate and loss rate can be rewritten as *ps* and (1–*p*)*s*, respectively (Figure 21.2). This is the binary case of a general time-reversible model (Tavaré 1986), which parameterizes changes between multiple states in a continuous-time Markov chain according to stationary probabilities of state presence and exchange rates or speeds of change between each pair of states. Figure 21.2 shows simulated trajectories of change under binary CTM processes with different speeds and stationary probabilities.

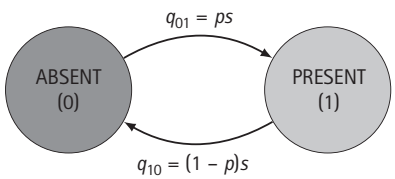


FIGURE 21.2 Binary continuous-time Markov process representing changes between presence and absence of a feature.

Transitions are annotated according to the gain rate (q_{01}) and loss rate (q_{10}) of the feature, with alternative parameterizations according to speed of change (s) and stationary probability of feature presence (p).

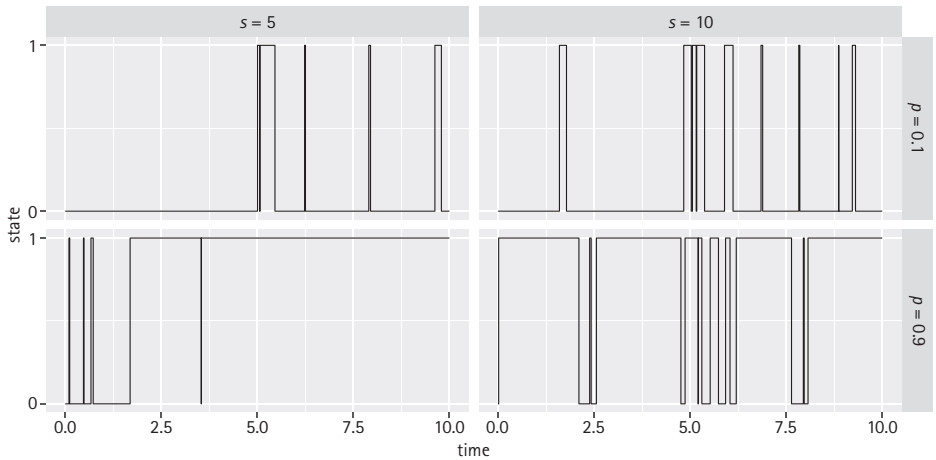


FIGURE 21.3 Simulated CTM processes showing transitions between states of a binary feature under different speeds ($s \in \{5, 10\}$) and stationary probabilities ($p \in \{0.1, 0.9\}$).

Classical CTM models make global estimates regarding the dynamics of evolution of a feature, but they do not always deal with the impact of population history and the variation in dynamics that external factors may bring about. Some linguistic work (Greenhill et al. 2017) has made use of modelling techniques that allow transition rates to vary across features and branch levels (Huelsenbeck and Suchard 2007), but this allows only post-hoc interpretations regarding which branches in a phylogeny are most affected by factors such as language contact, population density, etc. We expand this variable-rates (or 'heterotachy') approach to allow for the regressing of various

9780192886491_Book.indb 403 27-Mar-25 23:04:25

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

predictors, extralinguistic and linguistic, on the evolutionary parameters of the CTM process, profiting from advances made in biology and anthropology in assessing correlations between different co-evolving variables (see Lartillot and Poujol 2011; Lartillot 2013; Horvilleur and Lartillot 2014; Ross et al. 2016; Ringen et al. 2021; and Cathcart 2023 for a review).

We furthermore extend the models to allow both the speed of change and the stationary probability of a feature to vary according to one or more predictors. We do this in analogy with a recently proposed innovation in hierarchical Bayesian modelling, namely so-called distributional models, which allow both the predicted measure of central tendency and the dispersion of a response in a regression model to vary as a function of predictor variables (Rigby and Stasinopoulos 2005; Bürkner 2017).

A more refined understanding of whether certain extralinguistic predictors of change affect speed vs. stationary biases towards a particular feature is crucial for understanding linguistic evolution, in particular if the goal is to characterize the intrinsic dynamic of language faculty independently of the vagaries of population history. For instance, it helps answer questions on how complexity evolves, for example in morphology ('polysynthesis'). While the decrease of complexity has been successfully linked to contact and admixture of second language learners (Lupyan and Dale 2010; Bentz and Winter 2013; Widmer et al. 2021), the apparent increase of complexity in regions of greater social isolation is unexplained. Is this increase driven by a stationary long-term preference in the language faculty, or is it a side effect from more rapid change under isolation that lets grammaticalization accumulate in repeated waves?

Predictors can be either static or evolve with the response, they can be discrete or continuous, and, unlike classical models of correlated evolution (Pagel 1994), the design is not limited to evolving feature pairs. Cathcart et al. (2022) use this framework to investigate whether the evolution of irregular stem alternation types in Romance verbs is a byproduct of a general instability in lower-frequency verbs, or driven by a long-term preference for regular paradigms in lower-frequency verbs and irregular ones in higher-frequency ones. This was tested by assessing the effect of frequency on both speed of pattern change as well as stationary probability (i.e. long-term preference) for irregular patterns. Results indicated an effect of frequency on long-term preference for irregular patterns, but not on speed of change, indicating that frequency does not play a role in a verb's stability of inflectional patterns, but instead impacts timeless biases towards or against irregular patterns.

21.5 A CASE STUDY

Here, we focus on the phylogenetic and geospatial dimensions of the development of sound systems, probing the extent to which dynamics of geographic dispersal can be taken as a proxy for variation in sociolinguistic milieux that would have different impacts on linguistic evolution. We chose, as an example, the Austronesian languages,

9780192886491_Book.indb 404 27-Mar-25 23:04:25

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

a family exhibiting considerable variation in terms of phonological inventories across vastly different language ecologies. Different regions of the Austronesian-speaking world exhibit disparities in the number of contrastive segments that the languages in them display, and have undergone different processes of language change during their history. The western Pacific, in particular archipelagos like Vanuatu, has been characterized by small languages in stable multilingual environments in which there are strong language ideologies and pressures towards schismogenesis and esoterogeny. Polynesia, on the other hand, is characterized by greater linguistic homogeneity and less pronounced contact between speakers of different languages.

Geographic dispersals have a number of potential impacts on linguistic change which are not fully understood and partly at odds with each other. Rapid dispersals have the potential to bring speakers into contact with, and absorb, populations speaking a different language, which can lead to rapid change; this may not have been the case in Austronesian, however, as rapid dispersals brought Austronesian speakers to uninhabited areas, and the western Pacific, a region of longer settlement, bears witness to scenarios in which languages remaining closer to the Austronesian homeland for longer periods of time work to make themselves maximally divergent, in particular in terms of lexical items and phonological characteristics (Wurm 1967; François 2010, 2011; Rangelov et al. 2023). Similarly, language families often exhibit higher amounts of diversity (i.e. larger numbers of languages) closer to their homelands (Sapir 1916), reflecting greater degrees of schismogenesis. By contrast, languages further from the homeland, spoken in the vicinity of fewer closely related languages, may not exhibit as many schismogenetic effects.

Assuming that sound systems are prime indices of group identity in dense linguistic ecological niches (Evans 2019), the question arises whether schismogenetic effects primarily affect speed of segment replacement or the (stationary) probabilities of segment presence, or both together. This question can only be addressed with a distributional phylogenetic model. Concretely, we ask the following: (i) do lineages closer to the homeland undergo faster speeds of change in segmental inventories? And (ii) do they display a longer-term preference for more contrastive and non-contrastive segments than lineages that migrated further away and are more isolated from other languages. Or do we find both effects together?

In what follows, we explore the direct impact of patterns of geographic dispersal on linguistic evolution, jointly estimating the parameters of a phylogeographic model of language diffusion and a CTM model of changes in phonological inventories.

21.5.1 Data

We use data from PHOIBLE v. 2.0 (Moran and McCloy 2019), a large cross-linguistic database of segmental inventories which contains language-level metadata such as geographic coordinates at which a language is spoken. The data in PHOIBLE generally follows the analytical decisions made in the source from which it is extracted, and

9780192886491_Book.indb 405 27-Mar-25 23:04:25

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

often contains more than one record of an inventory for a single language. We randomly sampled a record for languages where more than one record was available, and retained only segments found in the Cross-Linguistic Transcription Systems database (CLTS; Anderson et al. 2018) that were found more than once across all languages, in order to maximize cross-linguistic comparability.

We use the Austronesian phylogeny of Gray et al. (2009), which consists of a Bayesian tree sample of several hundred languages, to conduct our phylogenetic comparative analyses. Branch lengths were scaled to one time unit per millennium, so that a speed of, say, $s = .5$ means two ($1/s$) expected changes (gains or losses) per millennium. We discard languages in the tree not found in PHOIBLE. For languages in PHOIBLE but not in the tree (such as Cocos Malay), we use a semi-automated script in order to graft languages in PHOIBLE but not in the phylogeny onto the most appropriate loci in each tree in the tree sample. In total, our data consist of 192 binary segmental features in 110 languages.

21.5.2 Methods

We build on the relaxed random walk model (RRW) of phylogeography, under which geospatial diffusion takes place according to a process of Brownian motion, the scale of which exhibits branch-level variation: on some branches of a phylogeny, geographic dispersal will take place at faster rates than on others. This model and extensions serve as the standard for modelling linguistic migration in continuous space (Lemey et al. 2010; Gill et al. 2017), and accurately recapitulate language families' spreads when dispersal involves expansion from a given point of origin, but not necessarily when wholesale migration from the point of origin has taken place (Neureiter et al. 2021); accordingly, the RRW and its extensions may not be appropriate for all of the world's language families. The basic RRW employed in this chapter is not sensitive to environmental features in the way that more sophisticated discrete-space models are (Bouckaert et al. 2012, 2018; Koile et al. 2022). We choose this model to illustrate ease of implementabilty in standard statistical programming languages.

Concretely, the latitude and longitude (written ℓ for brevity) at the tips of a phylogeny are assumed to follow a multivariate normal distribution (cf. O'Meara et al. 2006):

$$\ell^t \sim \text{MultiNormal}(\ell^r, \Sigma)$$

where ℓ^t is the vector of observed values at the tips of the tree, and ℓ^r is a vector the same length as ℓ^t which repeats the value at the root, which is unobserved; Σ represents the phylogenetic covariance between tips in the phylogeny. Under time-homogeneous Brownian motion, the phylogenetic covariance between languages i and j, $\sum_{ij}$, is equal to the sum of the lengths of the branches leading from the root of phylogeny to the most recent common ancestor (MRCA) of the two languages, multiplied by a parameter σ representing the scale of the diffusion process. Under a RRW model, $\sum_{ij}$ is equal to

9780192886491_Book.indb 406 27-Mar-25 23:04:26

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

$\sigma\rho^{\top}\zeta_{\mathrm{MRCA}_{ij}}$, where σ is the global scale of the process' diffusion, $\boldsymbol{\rho}$ is a vector of scale multipliers for each branch in the phylogeny and $\zeta_{\mathrm{MRCA}_{ij}}$ is a vector that contains the lengths of the branches that intervene between the root of the phylogeny and the most recent common ancestor of nodes *i* and *j*, with zeros corresponding to branches that do not intervene. We assume that for each phylogeny, $\mathbf{lon}^t$ and $\mathbf{lat}^t$ are distributed as above, with separate scale parameters σ^{lon}, σ^{lat} and $\boldsymbol{\rho}^{\mathrm{ton}}$, $\boldsymbol{\rho}^{\mathrm{tat}}$ across both dimensions of diffusion. For simplicity, we do not model correlation between the two dimensions, and do not account for measurement error in longitude and latitude values recorded for languages. We transform longitude and latitude coordinates to an equal earth projection (Šavrič et al. 2019), which minimizes distortions in distance found at different latitude values and accounts for the fact that coordinates on either side of the prime meridian are relatively close to each other (despite having highly disparate longitude values).

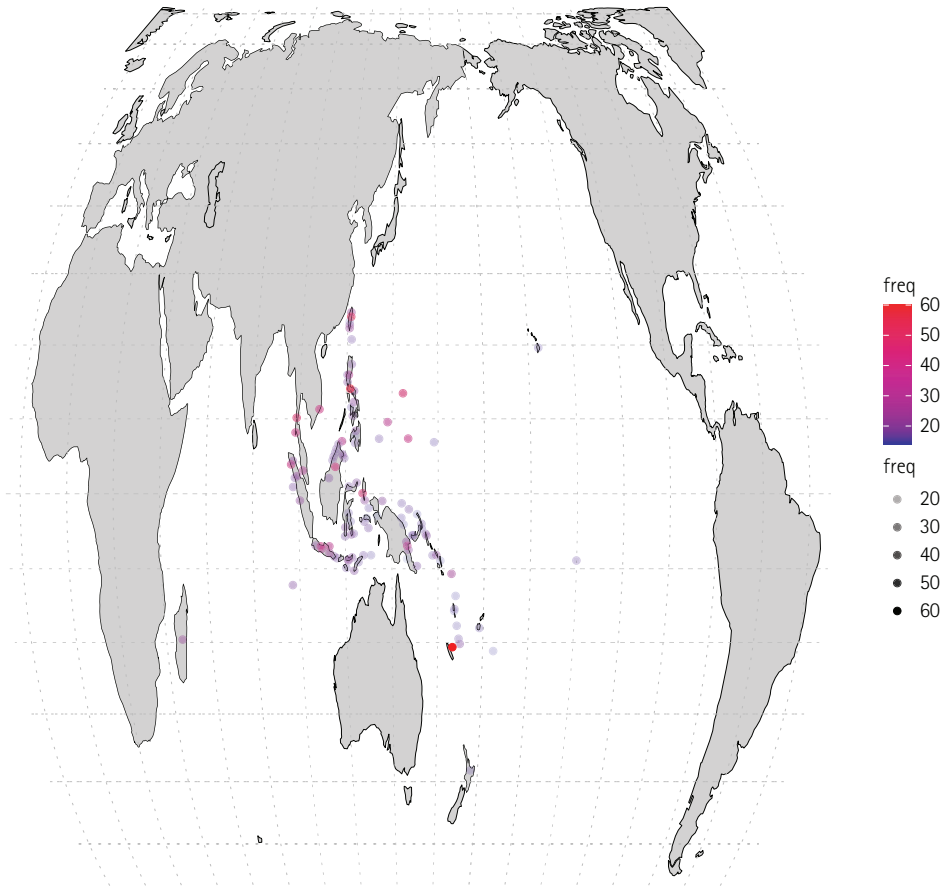


FIGURE 21.4 Austronesian languages in PHOIBLE. Colour and transparency indicate the number of segments in the inventory selected for the language.

9780192886491_Book.indb 407 27-Mar-25 23:04:28

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Using the parameters lon^r, lat^r, σ^{lon}, σ^{lat}, ρ^{lon}, ρ^{lat}, longitude and latitude values for internal nodes of the tree can be sampled from a conditional multivariate normal distribution. For a given sample, it is then possible to compute the great-circle distance between the root node of the tree and the descendant node of every branch in the tree, yielding values for branch-level displacements, $\delta_b : b \in \{1,\ldots,B\}$, i.e. the distance that the speech community represented by the descendant node has migrated from the linguistic homeland.

We link the quantity δ to the CTM process of segmental inventory evolution. For simplicity and tractability, we assume that each segment in each language can be gained or lost independently according to a binary CTM process. While this is not entirely justified given chain shifts and other dependencies, it has served an approximation with reasonable performance in previous work (e.g. Nikolaev and Grossman 2020; Guzmán Naranjo and Mertner 2022).

In line with our distributional modelling approach we parameterize sound change according to a speed of change and an expected stationary probability. On a given branch $b \in \{1,\ldots,B\}$, a segment $d \in \{1,\ldots,D\}$ evolves according to a binary CTM process with speed $s_{d,b}\tau$ (τ being the global speed of change in the tree) and stationary probability $p_{d,b}$. We force the speed multiplier $s_{d,b}$ and stationary probability $p_{d,b}$ into the interval (0,1) by using a logit transformation and assume the following form in a hierarchical (multi-level) model:

$$\text{logits}_{d,b} = \alpha^s_0 + \alpha^s_d + (\beta^S_0 + \beta^s_d)\log\delta_b + \gamma^s_{b,0}\sigma^s_d + \gamma^s_{b,d}\sigma^s_d$$

$$\text{logit}p_{d,b} = \alpha^p_0 + \alpha^p_d + (\beta^p_0 + \beta^p_d)\log\delta_b + \gamma^p_{b,0}\sigma^p_d + \gamma^p_{b,d}\sigma^p_d$$

In plain terms, the logit of the speed multiplier for segment d on branch b of the phylogeny is a function of an intercept-like baseline (α^s_0); a deflection to this baseline for segment d (α^s_d); coefficients representing the effects of displacement in each branch b (δ_b) on speed (β^s_0) and a deflection to this coefficient for segment d (β^s_d); as well as coefficients representing variation in branch-level speeds not explained by fluctuations in dispersal values, globally ($\gamma^s_{b,0}$) and at the segment level ($\gamma^s_{b,d}$); σ^s represents scale parameters allowing these coefficients to capture as much variance as is justified. The parameters pertaining to the stationary probability of segment d's presence for branch b have the same interpretation. The likelihood of the observed distributions of each segment under the model parameters and a phylogeny can be computed via Felsenstein's pruning algorithm (Figure 21.1).

The model described above can be fitted in a variety of ways. One possibility is to jointly infer all parameters, phylogeographic and phylogenetic, in one model, which can be computationally costly. Another option is to first infer the phylogeographic parameters, most importantly δ, separate from the phylogenetic parameters. Subsequently, posterior estimates of this quantity can be used as an input to the phylogenetic model described above as a variable with measurement error (Bürkner 2017; Driver et al. 2017; Bürkner et al. 2019). This is justified for models which posit the

9780192886491_Book.indb 408 27-Mar-25 23:04:33

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

directed influence of phylogeography on linguistic evolution and not vice versa. Here, we infer all parameters jointly, and leave a comparison of the merits of different methods for future work.

We place Normal (0,1) priors on the parameters α,β, and γ, HalfNormal (0,1) priors on the parameters ρ and σ, and a Uniform (0,20) prior on τ, which ensures that changes will not occur more frequently than once per fifty years (1/20 per millennium), roughly, over a small number of generations. To incorporate phylogenetic uncertainty, we fit the model on twenty-five trees from the tree sample and concatenate the resulting posterior samples. We use RStan (Carpenter et al. 2017) for model fitting and make all data and code available at https://github.com/chundrac/disp-phon).

21.5.3 Results

Figure 21.5 provides a visualization of dispersal trajectories inferred by our model; blue branches are closer to the inferred homeland. Regarding the accuracy of the inferred homeland, there is perhaps room for improvement: while Taiwan as a homeland is supported by this posterior distribution, the northern Philippines are as well, and this latter homeland is not in line with received wisdom. It is possible that the relatively small number of Austronesian taxa in PHOIBLE is responsible for this relatively high uncertainty, and that more precise results will be reached with a greater number of languages. Our model allocates credibility to potential homeland locations over water. This is behaviour potentially displayed by the RRW and other Gaussian process-based geospatial models. Implementing spatial constraints directly within programming languages like Stan (i.e. assigning zero prior probability to homeland locations that do not overlap with landmasses) is not straightforward, and at this point in time we are not fully aware of the practical issues surrounding the implementation of discrete, feature-aware models of migration (in general, these involve massive transition matrices between points on the globe, and dynamic solutions for storing pre-computed values). One post-hoc solution to this issue (which we do not employ here) is to discard samples after inference for which the root location is over water, which effectively produces results that would be obtained with a spatially informed prior.

Figure 21.6 shows the 95 per cent and 85 per cent posterior credible intervals (CIs) for β_0^s and β_0^p, model parameters representing the overall effect of dispersal on speed of change and long-term preference for segments. A commonly used criterion holds that distributions where the 95 per cent CI (i.e. the interval spanning from the 2.5 per cent to 97.5 per cent percentiles) does not contain zero, represent decisive evidence for an effect of a predictor on a response (Kruschke 2021). The results answer question (i) negatively: we find a decisive positive effect of dispersal distance on speed of change, i.e. slower, rather than faster, change closer to the homeland. But they answer question (ii) in the affirmative: we find a decisive negative effect of dispersal distance on stationary probability of segment presence, i.e. languages show more segments closer to the

9780192886491_Book.indb 409 27-Mar-25 23:04:33

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

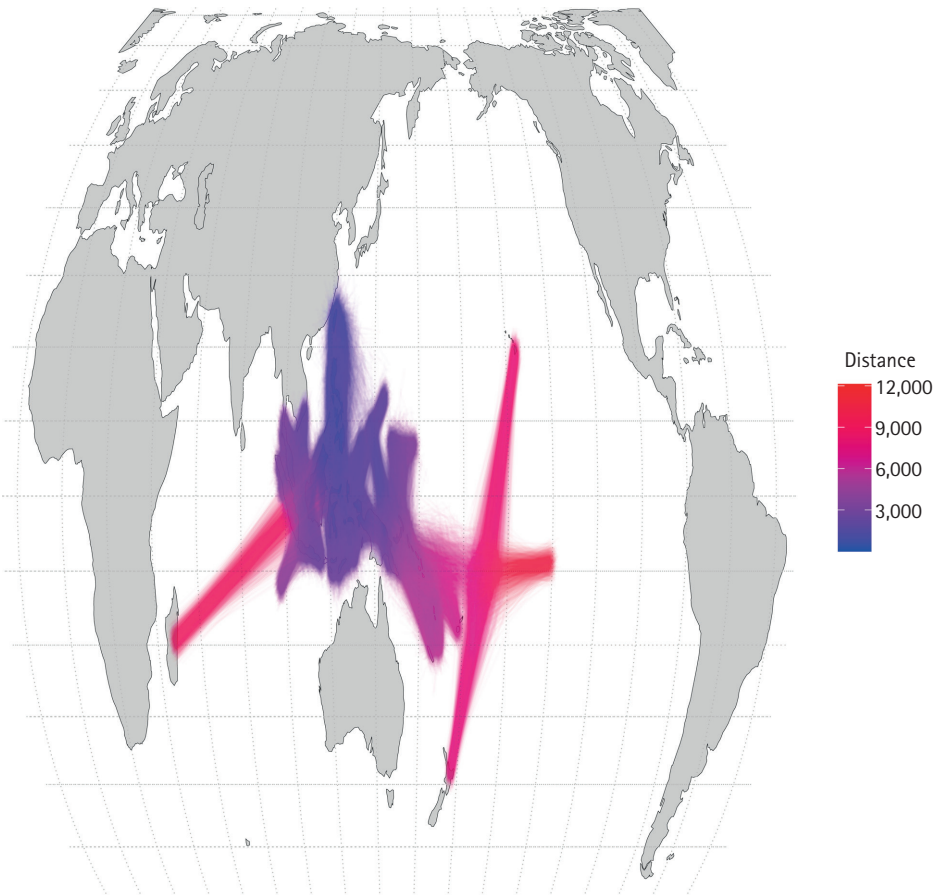


FIGURE 21.5 Posterior samples of phylogeographic trajectories inferred by our model. Colours represent distances (in km) from root location of tree for child of each branch.

homeland. This indicates that branches undergoing more extensive migration undergo faster speeds of phonemic change, but display a long-term trend towards having fewer consonants than branches closer to the point of origin of the family.

To further understand the reasons for this result, we inspect the branch-level phylogeographic diffusion patterns for one of the twenty-five tree samples over which we run our model. Figure 21.7 colours branches of the tree according to the distance between the root of the tree and the child nodes of each branch. Blue branches have undergone less displacement. Polynesian languages (at the top of the tree) are conspicuous in having smaller inventories and being geographically distant from the Austronesian homeland. Larger inventories can be found in languages that are closer to the homeland (e.g. in the bottom half of the tree). Our model may have inferred that

9780192886491_Book.indb 410 27-Mar-25 23:04:33

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

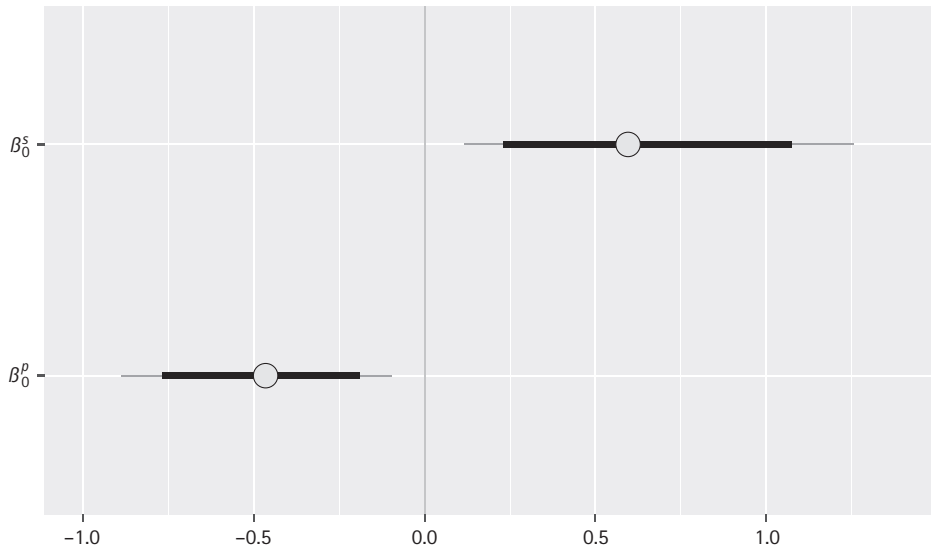


FIGURE 21.6 95% and 85% posterior credible intervals (CIs) for β_0^S and β_0^p. Distributions where the 95% CI does not contain zero represent decisive evidence for an effect of a predictor on a response.

these larger inventories are at least partially the product of shared retentions preserved by lower speeds of change, rather than brought on by rapid complexification of inventories due to extreme schismogenetic pressures.

Inspection of additional model parameters also allows us to assess the degree to which dynamics of change vary at the segment level. Figure 21.8 shows coefficients representing segment-level deflections from the overall effect of diffusion rate on speed of change and stationary probability of presence. The left-hand graphic shows 85 per cent and 95 per cent CIs for coefficients reflecting variation in the effect of dispersal distance on speed of change. For visibility, we show only segments with the thirty lowest and thirty highest median values. Here, all 95 per cent CIs (and most 85 per cent CIs) overlap with zero, indicating that there are not many decisive segment-level deviations from the overall effect of dispersal on speed of change. The right-hand graph arranges segments according to the degree to which their long-term probability of presence is affected by dispersal distance (as above, with only the thirty lowest and thirty highest median values). Here, we see a larger number of 85 per cent and 95 per cent CIs that do not overlap with zero, indicating that segments vary considerably more in terms of their resistance to the effects of dispersal. At the top of the graph, we see segments for which we expect a decisively lower-than-baseline stationary probability of presence as distance from the homeland increases; the segments /z/, /q/, and /ħ/ are restricted to languages spoken on Taiwan, close to the inferred homeland. At the other end of the spectrum is /m/, found in all languages; the stationary probability of /m/ will be higher than expected for increases in distance from the homeland, as

9780192886491_Book.indb 411 27-Mar-25 23:04:34

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

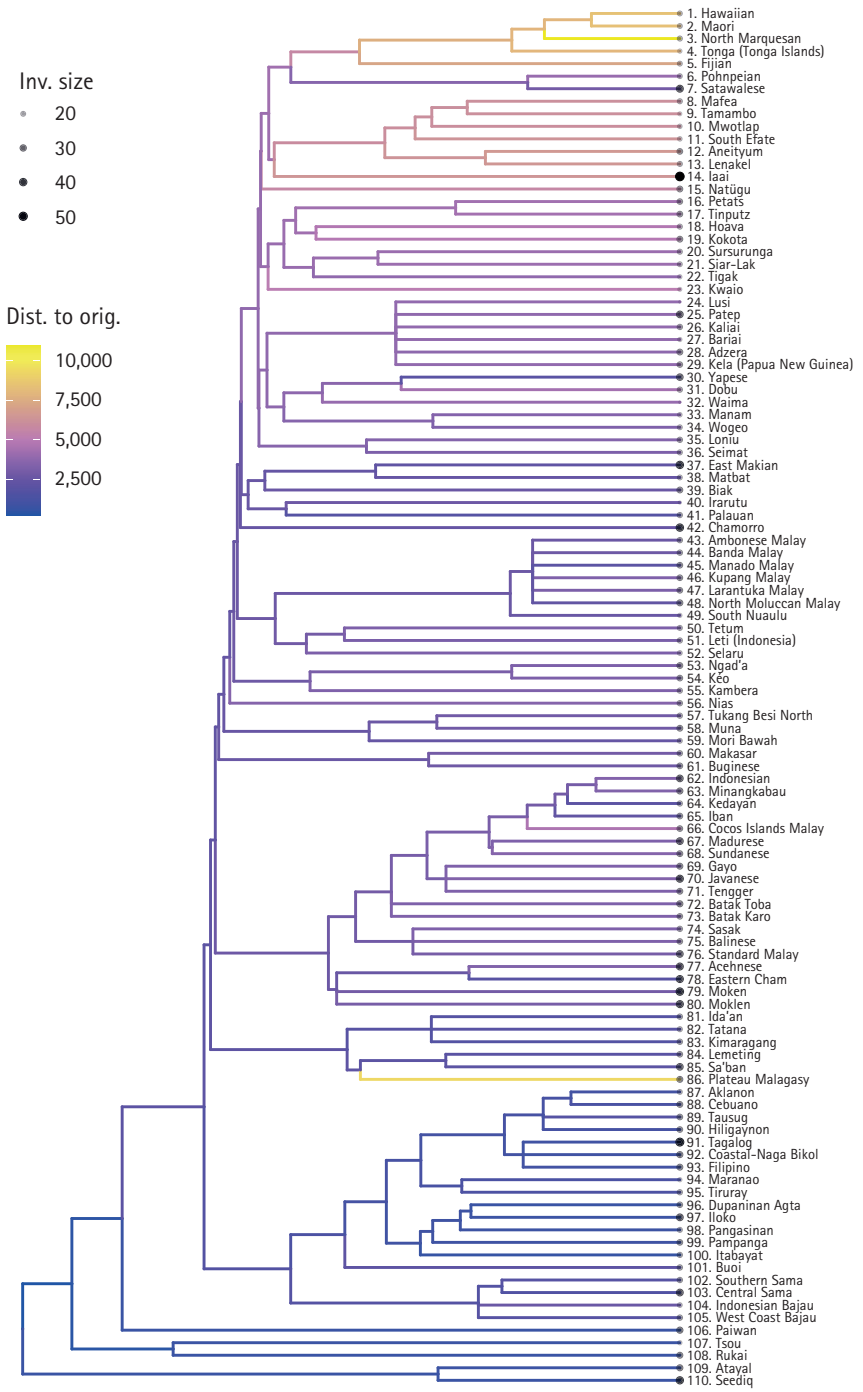


FIGURE 21.7 Median branch-level dispersals (great-circle distances in km between the inferred root location and the inferred/observed locations of the child nodes of each branch) for a tree from the tree sample, visualized using ggtree (Yu et al. 2018).

Blue branches have undergone less dispersal than yellow ones. Dot sizes at tips of the tree represent inventory sizes for each language.

9780192886491_Book.indb 412 27-Mar-25 23:04:34

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

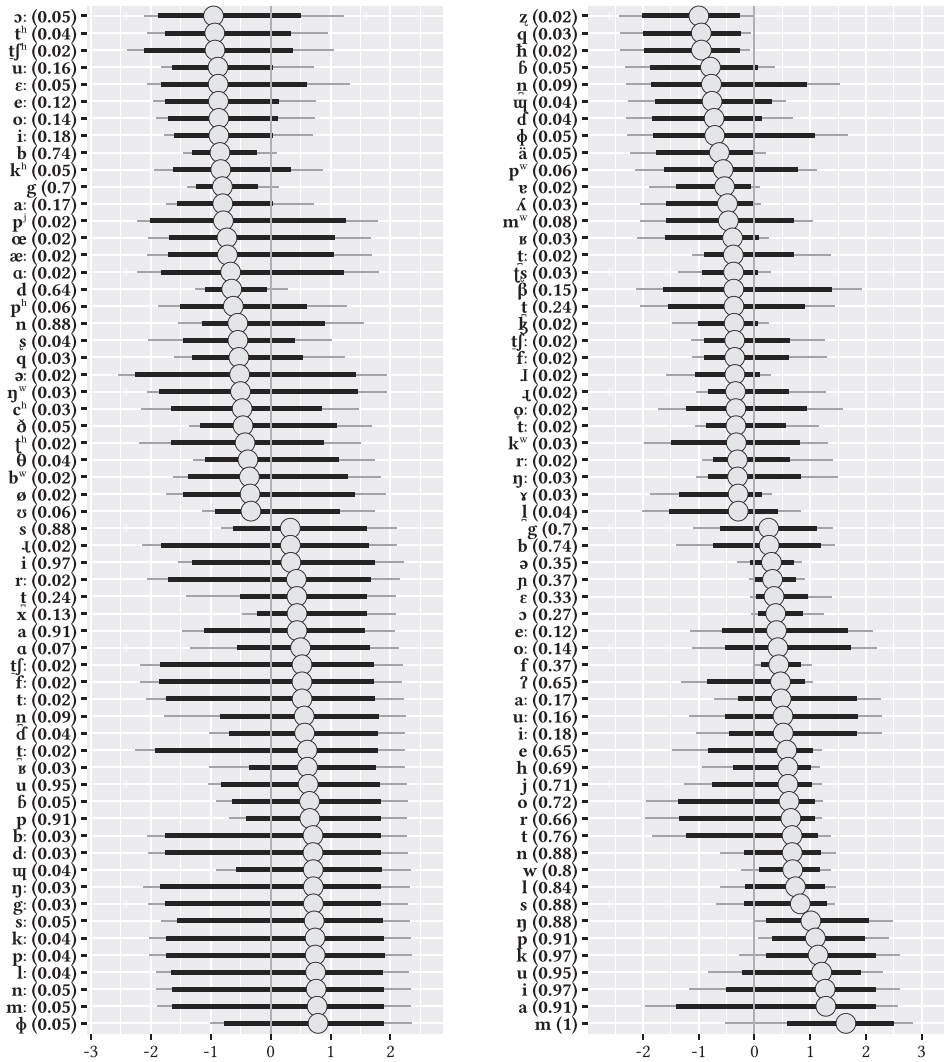


FIGURE 21.8 95% and 85% posterior credible intervals for segment-level deflections to the global effect of dispersal on speed of change (left) and stationary probability of presence (right).
The relative frequency of each segment in the dataset is given in parentheses.

it is a segment universally preferred. In general, segments making up a core of most languages' inventories appear to be less sensitive to the effects of dispersal. These include nasals like /m/ as well as relatively basic segments like /k/, /b/, /d/, etc. In general, the sounds most resistant to changes coinciding with dispersal tend to include consonants that are thought to be more 'primal' and basic in inventories (Bybee and Easterday 2022).

9780192886491_Book.indb 413 27-Mar-25 23:04:35

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

21.5.4 Discussion

The coefficients for speed and stationary probability provide a way of assessing the degree to which a particular linguistic feature is resistant to the pressures of contact (if our model of phylogeographic dispersal can, in fact, be taken as a proxy for contact proneness). Some idiosyncrasies seen here may be artefacts of using PHOIBLE's segmental representations directly. An alternative is to use features (e.g. labial, nasal, etc.), if care is taken to code the data such that interactions of interest (e.g. velar plosive) are taken into account.

Given the large amount of literature arguing for the effect of long standing stable multilingualism on patterns of schismogenesis, it is surprising to see that lineages that undergo more geographic dispersal, rather than those with longer histories of local settlement, show higher rates of phonemic change. Further inspection of these rates showed that branches exhibiting greater dispersals tend to lead to languages varying considerably in the size of their inventories, explaining why we see this effect at a large family-wide scale.

At this stage, we have only tentative explanations for this result. As mentioned previously, in any other scenario, we might interpret rapid language change accompanying large-scale migration as a signature of contact-induced changes brought on by the absorption of other populations, but this does not hold for Austronesian, as longer migrations largely brought speakers to uninhabited locations. It is possible that the faster linguistic changes we detect coincided with a sort of societal complexification (perhaps involving dialect levelling) that accompanied the development of technological advances needed to carry out large-scale migrations. Padilla-Iglesias et al. (2020) link geographic isolation of the sort found in the eastern Pacific to acceleration of linguistic change, in the form of word loss.

Our results, obtained based on a wide phylogenetic scale, do not invalidate the idea that stable multilingualism fosters pressures to differentiate oneself from other social groups. It may be the case that using phylogeographic variables as a proxy for the social pressures of interest is not appropriate, and other variables need to be included. Recent attempts to characterize distributions of languages and linguistic features chiefly according to geographic predictors have produced mixed results (Urban 2021; Shcherbakova et al. 2023). In fact, a larger number of continuous co-evolving predictors representing demographic information may need to be integrated into such models (cf. Ringen et al. 2021). It may also be the case that using segmental inventories as the response variable of interest does not capture fine-grained patterns of schismogenesis. Finally, it could be that the relationship between phylogeographic diffusion and linguistic change varies at cladistic levels smaller than the family, such as within individual subgroups.

We hope that distributional phylogenetic modelling allows for tackling such questions with more specific and informative predictors, disentangling the various forces that shape the evolutionary dynamics of language.

9780192886491_Book.indb 414 27-Mar-25 23:04:35

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

21.6 Outlook

Bayesian models of linguistic evolution present a number of intriguing avenues for understanding the intrinsic dynamics of the language faculty. Considerable challenges lie ahead in terms of the integration of phylogenetic models, which incorporate flexible modelling of language change, with areal models showing an explicit representation of geospatial neighbourhoods. Here, we discussed some potential rapprochements between these two families of models, and presented the results of a case study assessing the effect of phylogeography on linguistic evolution, couched in the nascent framework of distributional phylogenetic modelling. A benefit of this approach is that it can help recover deep historical pressures that cannot be detected solely from synchronic distributions of linguistic and extralinguistic features. A careful inspection of model parameters shed light on why our result obtained, and we outline additional ways in which models of this sort can be expanded and bring more linguistic insights into consideration. Model expansion in this direction will be key to validating a number of hypotheses regarding linguistic evolution. Many tasks remain in fully understanding the forces that shape linguistic evolution. We hope that with an increase of flexible approaches like the one outlined here, accompanied by careful model criticism, we will move closer to this goal.

References

Abouheif, E. (1999). 'A method for testing the assumption of phylogenetic independence in comparative data', *Evolutionary Ecology Research* 1(8): 895–909.

Anderson, C., Tresoldi, T., Chacon, T., et al. (2018). 'A cross-linguistic database of phonetic transcription systems', in *Yearbook of the Poznan Linguistic Meeting*, Vol. 4, 21–53. Berlin: De Gruyter Open.

Atkinson, Q. D. (2011). 'Phonemic diversity supports a serial founder effect model of language expansion from Africa', *Science* 332(6027): 346–349.

Bailey, C.-J. N. (1973). *Variation and Linguistic Theory*. Washington, DC: Center for Applied Linguistics.

Barbieri, C., Blasi, D. E., Arango-Isaza, E., et al. (2022). 'A global analysis of matches and mismatches between human genetic and linguistic histories', *Proceedings of National Academy of Science* 119(47): e2122084119.

Beaulieu, J. M. and O'Meara, B. C. (2014). 'Hidden Markov models for studying the evolution of binary morphological characters', in L. Z. Garamszegi (ed.), *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology: Concepts and Practice*, 395–408. Heidelberg, New York, Dordrecht, London: Springer.

Bentz, C. and Winter, B. (2013). 'Languages with more second language learners tend to lose nominal case', *Language Dynamics and Change* 3: 1–27.

Bickel, B. (2007). 'Typology in the 21st century: Major current developments', *Linguistic Typology* 11: 239–251.

9780192886491_Book.indb 415 27-Mar-25 23:04:35

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Bickel, B. (2015). 'Distributional typology: Statistical inquiries into the dynamics of linguistic diversity', in B. Heine and H. Narrog (eds.), *The Oxford Handbook of Linguistic Analysis, 2nd Edition*, 901–923. Oxford: Oxford University Press.

Bickel, B. (2019). 'Large and ancient linguistic areas', in M. Crevels, J.-M. Hombert, and P. Muysken (eds.), *Language Dispersal, Diversification, and Contact: A Global Perspective*, 78–102. Oxford: Oxford University Press [pre-print available at http://www.ivs.uzh.ch/bickel-files/papers/Ancient_areasBickel2015.pdf].

Bickel, B., Witzlack-Makarevich, A., Choudhary, K. K., et al. (2015). 'The neurophysiology of language processing shapes the evolution of grammar: Evidence from case marking', *PLoS ONE* 10, e0132819.

Bickel, B., Giraud, A.-L., Zuberbühler, K., and Van Schaik, C. P. (2023). 'Language follows a unique mode of extra-genomic evolution', *Nature Human Behaviour* [under review].

Blasi, D. E., Moran, S., Moisik, S. R., et al. (2019). 'Human sound systems are shaped by post-Neolithic changes in bite configuration', *Science* 363(6432): 1–10.

Blei, D. M., Ng, A. Y., and Jordan M. I. (2003). 'Latent Dirichlet allocation', *Journal of Machine Learning Research* 3: 993–1022.

Blomberg, S. P., Garland Jr., T., and Ives, A. R. (2003). 'Testing for phylogenetic signal in comparative data: Behavioral traits are more labile', *Evolution* 57(4): 717–45.

Bollback, J. P. (2006). 'SIMMAP: Stochastic character mapping of discrete traits on phylogenies', *BMC Bioinformatics* 7: 88.

Borges, R., Machado, J. P., Gomes, C., et al. (2019). 'Measuring phylogenetic signal between categorical traits and phylogenies', *Bioinformatics* 35(11): 1862–1869.

Bouckaert, R., Lemey, P., Dunn, M., et al. (2012). 'Mapping the origins and expansion of the Indo-European language family', *Science* 337(6097): 957–960.

Bouckaert, R. R., Bowern, C., and Atkinson, Q. D. (2018). 'The origin and expansion of Pama-Nyungan languages across Australia', *Nature Ecology & Evolution* 2(4): 741–749.

Bouckaert, R., D. Redding, O. Sheehan, T., et al. (2022). 'Global language diversification is linked to socio-ecology and threat status', *SocArXiv*. July 20, https://osf.io/f8tr6/download. Accessed 10 June 2023.

Bowern, C. (2012). 'The riddle of Tasmanian languages', *Proceedings of the Royal Society B: Biological Sciences* 279(1747): 4590–4595.

Bürkner, P.-C. (2017). 'Bayesian distributional non-linear multilevel modeling with the r package brms', *arXiv preprint arXiv:1705.11123*.

Bürkner, P.-C., Schulte, N., and Holling, H. (2019). 'On the statistical and practical limitations of Thurstonian IRT models', *Educational and Psychological Measurement* 79(5) 827–854.

Bybee, J. and Easterday, S. (2022). 'Primal consonants and the evolution of consonant inventories', *Language Dynamics and Change* 13: 1–33.

Bürkner, P.-C. (2017). 'brms: An R package for Bayesian multilevel models using Stan', *Journal of Statistical Software* 80(1): 1–28.

Carling, G. and Cathcart, C. (2021a). 'Evolutionary dynamics of Indo-European alignment patterns', *Diachronica* 38: 358–412.

Carling, G. and Cathcart, C. (2021b). 'Reconstructing the evolution of Indo-European Grammar', *Language* 97: 1–38.

Carpenter, B., Gelman, A., Hoffman, M. D., et al. (2017). 'Stan: A probabilistic programming language', *Journal of Statistical Software* 76(1): 1–32.

9780192886491_Book.indb 416 27-Mar-25 23:04:35

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Cathcart, C. (2023). 'Rate variation in language change: Toward distributional phylogenetic modeling', in A. Karakostis and G. Jäger (eds.), *Biocultural Implications: An Agenda for Integrative Approaches*, 179–202. Tübingen: Kerns.

Cathcart, C., Herce, B., and Bickel, B. (2022). 'Decoupling speed of change and long-term preference in language evolution: Insights from romance verb stem alternations', in Ravignani, A. et al. (eds.), *The Evolution of Language: Proceedings of the Joint Conference on Language Evolution (JCoLE)*, 101–108. Nijmegen, https://evolang.org/jcole2022/proceedings/jcole2022_proceedings.pdf.

Cathcart, C. A. (2020). 'A probabilistic assessment of the Indo-Aryan Inner-Outer hypothesis', *Journal of Historical Linguistics* 10: 42–86.

Cathcart, C. A. (2022). 'Dialectal layers in West Iranian: A hierarchical Dirichlet process approach to linguistic relationships', *Transactions of the Philological Society* 120(1): 1–31.

Cathcart, C. A., Hölzl, A., Jäger, G., et al. (2020). 'Numeral classifiers and number marking in Indo-Iranian: A phylogenetic approach', *Language Dynamics and Change* 1(aop): 1–53.

Chang, W. and Michael, L. (2014). 'A relaxed admixture model of language contact', *Language Dynamics and Change* 4(1): 1–26.

Chang, W., Cathcart, C., Hall, D., and Garrett, A. (2015). 'Ancestry-constrained phylogenetic analysis supports the Indo-European Steppe Hypothesis', *Language* 91(1): 194–244.

Cysouw, M. (2011). 'Understanding transition probabilities', *Linguistic Typology* 15: 415–431.

Daumé III, H. (2009). 'Non-parametric Bayesian areal linguistics', in *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL*, Boulder, CO, 593–601. Association for Computational Linguistics.

Driver, C. C., Oud, J. H., and Voelkle, M. C. (2017). 'Continuous time structural equation modeling with R package CTSEM', *Journal of Statistical Software* 77: 1–35.

Dryer, M. S. (1989). 'Large linguistic areas and language sampling', *Studies in Language* 13(2): 257–292.

Dunn, M., Greenhill, S. J., Levinson, S. C., and Gray, R. D. (2011). 'Evolved structure of language shows lineage-specific trends in word-order universals', *Nature* 473(7345): 79–82.

Dunn, M., Dewey, T. K., Arnett, C., et al. (2017). 'Dative sickness: A phylogenetic analysis of argument structure evolution in Germanic', *Language* 93(1): e1–e22.

Evans, N. (2019). 'Linguistic divergence under contact', in *Historical Linguistics* 2015, 564–591. Amsterdam, New York: John Benjamins.

Everett, C. (2013). 'Evidence for direct geographic influences on linguistic sounds: The case of ejectives', *PloS one* 8(6): e65275.

Everett, C., Blasi, D. E., and Roberts, S. G. (2015). 'Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots', *Proceedings of the National Academy of Sciences* 112(5): 1322–1327.

Felsenstein, J. (1981). 'Evolutionary trees from DNA sequences: A maximum likelihood approach', *Journal of Molecular Evolution* 17(6): 367–76.

Felsenstein, J. (2004). *Inferring Phylogenies*. Sunderland, MA: Sinauer Associates.

FitzJohn, R. G., Maddison, W. P., and Otto, S. P. (2009). 'Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies', *Systematic Biology* 58(6): 595–611.

François, A. (2010). 'The languages of Vanikoro: Three lexicons and one grammar', in B. Evans (ed.), *Discovering History Through Language: Papers in honour of Malcolm Ross*, Volume 605 of *Pacific Linguistics*. Canberra: Research School of Pacific and Asian Studies, Australian National University.

9780192886491_Book.indb 417 27-Mar-25 23:04:35

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

François, A. (2011). 'Social ecology and language history in the northern Vanuatu linkage: A tale of divergence and convergence', *Journal of Historical Linguistics* 1(2): 175–246.
Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.
Gelman, A., Carlin, J. B., Stern, H. S., et al. (2013). *Bayesian Data Analysis*. Boca Raton, FL: CRC press.
Geman, S. and Geman, D. (1984). 'Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images', *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6: 721–741.
Gill, M. S., Tung Ho, L. S., Baele, G., et al. (2017). 'A relaxed directional random walk model for phylogenetic trait evolution', *Systematic Biology* 66(3): 299–319.
Gray, R. D., Drummond, A. J., and Greenhill, S. J. (2009). 'Language phylogenies reveal expansion pulses and pauses in pacific settlement', *Science* 323(5913): 479–483.
Greenberg, J. H., Osgood, C. E., and Saporta, S. (1954). 'Language change', In C. Osgood and T. Sebeok (eds.), *Psycholinguistics: A Survey of Theory and Research Problems*, 146–63. Baltimore: Waverly Press.
Greenhill, S. J., Wu, C.-H., Hua, X., et al. (2017). 'Evolutionary dynamics of language systems', *Proceedings of the National Academy of Sciences* 114(42): E8822–E8829.
Greenhill, S. J., Heggarty, P., and Gray, R. D. (2021). 'Bayesian phylolinguistics', in C. Bowern and B. Evans (eds.), *The Handbook of Historical Linguistics*, 226–253. London: Wiley-Blackwell.
Guzmán Naranjo, M. and Becker, L. (2022). 'Statistical bias control in typology', *Linguistic Typology* 26(3): 605–670.
Guzmán Naranjo, M. and Mertner, M. (2022). 'Estimating areal effects in typology: A case study of African phoneme inventories', *Linguistic Typology* 27(2): 455–480.
Hahn, M. and Xu, Y. (2022). 'Crosslinguistic word order variation reflects evolutionary pressures of dependency and information locality', *Proceedings of the National Academy of Sciences* 119(24): e2122604119.
Harmon, L. J. (2019). *Phylogenetic Comparative Methods*. GitHub, https://lukejharmon.github.io/pcm/.
Heggarty, P., Anderson, C., Scarborough, M., et al. (2023). 'Language trees with sampled ancestors support a hybrid model for the origin of Indo-European languages', *Science* 381(6656): eabg0818.
Horvilleur, B. and Lartillot, N. (2014). 'Monte Carlo algorithms for Brownian phylogenetic models', *Bioinformatics* 30(21): 3020–3028.
Huelsenbeck, J. P. and Suchard, M. A. (2007). 'A nonparametric method for accommodating and testing across-site rate variation', *Systematic Biology* 56(6): 975–987.
Huelsenbeck, J. P., Nielsen, R., and Bollback, J. P. (2003). 'Stochastic mapping of morphological characters', *Systematic Biology* 52: 131–158.
Ivani, J. and Bickel, B. (2024). 'Databases for comparative syntactic research', in S. Barbiers, N. Corver, and M. Polinsky (eds.), *The Cambridge Handbook of Comparative Syntax*. Cambridge: Cambridge University Press.
Jaeger, T. F., Graff, P., Croft, W., and Pontillo, D. (2011). 'Mixed effect models for genetic and areal dependencies in linguistic typology', *Linguistic Typology* 15: 281–320.
Jing, Y., Widmer, P., and Bickel, B. (2023). 'Word order evolves at similar rates in main and subordinate clauses', *Diachronica* 40: 1–32.
Jäger, G. and Wahle, J. (2021). 'Phylogenetic typology', *Frontiers in Psychology* 12: 1–15.
Kelly, L. J. and Nicholls, G. K. (2017). 'Lateral transfer in Stochastic Dollo models', *The Annals of Applied Statistics* 11(2): 1146–1168.

9780192886491_Book.indb 418 27-Mar-25 23:04:35

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Kirby, S. (2017). 'Culture and biology in the origins of linguistic structure', *Psychonomic Bulletin & Review* 24: 118–137.

Koile, E., Greenhill, S. J., Blasi, D. E., et al. (2022). 'Phylogeographic analysis of the Bantu language expansion supports a rainforest route', *Proceedings of the National Academy of Sciences* 119(32): e2112853119.

Kruschke, J. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, Jags, and Stan*. London: Academic Press.

Kruschke, J. K. (2021). 'Bayesian analysis reporting guidelines', *Nature Human Behaviour* 5(10): 1282–1291.

Lartillot, N. (2013). 'Phylogenetic patterns of GC-biased gene conversion in placental mammals and the evolutionary dynamics of recombination landscapes', *Molecular Biology and Evolution* 30(3): 489–502.

Lartillot, N. and Poujol, R. (2011). 'A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters', *Molecular Biology and Evolution* 28(1): 729–744.

Lemey, P., Rambaut, A., Welch, J. J., and Suchard, M. A. (2010). 'Phylogeography takes a relaxed random walk in continuous space and time', *Molecular Biology and Evolution* 27(8): 1877–1885.

Lupyan, G. and Dale, R. (2010). 'Language structure is partly determined by social structure', *PLoS ONE* 5: e8559.

Macklin-Cordes, J. L., Bowern, C., and Round, E. R. (2021). 'Phylogenetic signal in phonotactics', *Diachronica* 38(2): 210–258.

Maddison, W. P., Midford, P. E., and Otto, S. P. (2007). 'Estimating a binary character's effect on speciation and extinction', *Systematic Biology* 56(5): 701–710.

Maddieson, I., Flavier, S., Marsico, E., et al. (2013). 'LAPSyd: Lyon-Albuquerque phonological systems database', in *Proceedings of 14th Interspeech Conference*, Lyon, 25–29 August 2013, 3022–3026.

Maslova, E. (2000). 'A dynamic approach to the verification of distributional universals', *Linguistic Typology* 4: 307–333.

Maurits, L. and Griffiths, T. L. (2014). 'Tracing the roots of syntax with Bayesian phylogenetics', *Proceedings of the National Academy of Sciences* 111(37)', 13576–13581.

McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Boca Raton, FL: CRC Press.

Michael, L., Chang, W., and Stark, T. (2014). 'Exploring phonological areality in the circum-Andean region using a naive Bayes classifier', *Language Dynamics and Change* 4(1): 27–86.

Mimno, D., Wallach, H., and McCallum, A. (2008). 'Gibbs sampling for logistic normal topic models with graph-based priors', in *NIPS Workshop on Analyzing Graphs*, Volume 61.

Moran, S. and McCloy, D. (eds.) (2019). *PHOIBLE 2.0*. Jena: Max Planck Institute for the Science of Human History.

Murawaki, Y. and Yamauchi, K. (2018). 'A statistical model for the joint inference of vertical stability and horizontal diffusibility of typological features', *Journal of Language Evolution* 3(1): 13–25.

Narroll, R. (1961). 'Two solutions to Galton's Problem', *Philosophy of Science* 28: 15–29.

Nettle, D. (1999). 'Is the rate of linguistic change constant?', *Lingua* 108(2-3): 119–136.

Neureiter, N., Ranacher, P., Efrat-Kowalsky, N., et al. (2022). 'Detecting contact in language trees: A Bayesian phylogenetic model with horizontal transfer', *Humanities and Social Sciences Communications* 9(1): 1–14.

9780192886491_Book.indb 419 27-Mar-25 23:04:35

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Neureiter, N., Ranacher, P., van Gijn, R., et al. (2021). 'Can Bayesian phylogeography reconstruct migrations and expansions in linguistic evolution?', *Royal Society Open Science* 8(1): 201079.
Nikolaev, D. and Grossman, E. (2020). 'Consonant co-occurrence classes and the feature-economy principle', *Phonology* 37(3): 419–451.
Niyogi, P., Berwick, R. C., et al. (1997). 'A dynamical systems model for language change', *Complex Systems* 11(3): 161–204.
O'Meara, B. C., Ané, C., Sanderson, M. J., and Wainwright, P. C. (2006). 'Testing for different rates of continuous trait evolution using likelihood', *Evolution* 60(5): 922–933.
Padilla-Iglesias, C., Gjesfjeld, E., and Vinicius, L. (2020). 'Geographical and social isolation drive the evolution of Austronesian languages', *PloS one* 15(12): e0243171.
Pagel, M. (1994). 'Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters', *Proceedings of the Royal Society of London B* 255: 37–45.
Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). 'Inference of population structure using multilocus genotype data', *Genetics* 155(2): 945–959.
Ranacher, P., Neureiter, N., van Gijn, R., et al. (2021). 'Contact-tracing in cultural evolution: A Bayesian mixture model to detect geographic areas of language contact', *Journal of The Royal Society Interface* 18(181): 20201031.
Rangelov, T., Walworth, M., and Barbour, J. (2023). 'A multifaceted approach to understanding unexpected sound change: The bilabial trills of Vanuatu's Malekula island', *Diachronica* 40: 384–432.
Reesink, G., Singer, R., and Dunn, M. (2009). 'Explaining the linguistic diversity of Sahul using population models', *PLoS Biology* 7: e1000241.
Revell, L. J. (2013). 'Two new graphical methods for mapping trait evolution on phylogenies', *Methods in Ecology and Evolution* 4: 754–759.
Rigby, R. A. and D. M. Stasinopoulos (2005). 'Generalized additive models for location, scale and shape', *Journal of the Royal Statistical Society Series C: Applied Statistics* 54(3): 507–554.
Ringen, E., Martin, J. S., and Jaeggi, A. (2021). 'Novel phylogenetic methods reveal that resource-use intensification drives the evolution of "complex" societies', https://ecoevorxiv.org/repository/object/4119/download/8181/. Accessed 1 Mar 2022.
Robert, C. P. and Casella, G. (2010). 'Metropolis–hastings algorithms', in C. Robert and G. Casella (eds.), *Introducing Monte Carlo Methods with R*, 167–197. New York: Springer.
Roberts, S. G. (2018). 'Robust, causal, and incremental approaches to investigating linguistic adaptation', *Frontiers in Psychology* 9: 166.
Ross, C. T., Strimling, P., Ericksen, K. P., et al. (2016). 'The origins and maintenance of female genital modification across Africa: Bayesian phylogenetic modeling of cultural evolution under the influence of selection', *Human Nature* 27: 173–200.
Sapir, E. (1916). *Time Perspective in Aboriginal American Culture, a Study in Method*. Ottawa, CA: Government Printing Bureau.
Šavrič, B., Patterson, T., and Jenny, B. (2019). 'The equal earth map projection', *International Journal of Geographical Information Science* 33(3): 454–465.
Shcherbakova, O., Michaelis, S. M., Haynie, H. J., et al. (2023). 'Societies of strangers do not speak less complex languages', *Science Advances* 9(33): eadf7704.
Shirtz, S., Talamo, L., and Verkerk, A. (2021). 'The evolutionary dynamics of negative existentials in Indo-European', *Frontiers in Communication* 6: 661862, doi: 10.3389/fcomm.
Shringarpure, S. and Xing, E. P. (2009). 'mStruct: Inference of population structure in light of both genetic admixing and allele mutations', *Genetics* 182(2): 575–593.

9780192886491_Book.indb 420 27-Mar-25 23:04:35

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Sisson, S. A. and Fan, Y. (2011). 'Likelihood-free MCMC', *Handbook of Markov Chain Monte Carlo*: 313–335.

Stone, J. V. (2013). *Bayes' Rule: A Tutorial Introduction to Bayesian Analysis*. Warsaw: Sebtel Press.

Syrjänen, K., Honkola, T., Lehtinen, J., et al. (2016). 'Applying population genetic approaches within languages: Finnish dialects as linguistic populations', *Language Dynamics and Change* 6: 235–283.

Tavaré, S. (1986). 'Some probabilistic and statistical problems in the analysis of DNA sequences', *Lectures on Mathematics in the Life Sciences* 17(2): 57–86.

Urban, M. (2021). 'The geography and development of language isolates', *Royal Society Open Science* 8(4): 202232.

Wichmann, S., Holman, E. W., and Brown, C. H. (2018). The ASJP database (version 18), http://asjp.clld.org/.

Widmer, M., Auderset, S., Widmer, P., et al. (2017). 'NP recursion over time: Evidence from Indo-European', *Language* 93: 1–36.

Widmer, M., Jenny, M., Behr, W., and Bickel, B. (2021). 'Morphological structure can escape reduction effects from mass admixture of second language speakers: Evidence from Sino-Tibetan', *Studies in Language* 45: 707–752.

Witzlack-Makarevich, A., Zakharko, T., Bierkandt, L., et al. (2016). 'Decomposing hierarchical alignment: Co-arguments as conditions on alignment and the limits of referential hierarchies as explanations in verb agreement', *Linguistics* 54: 531–561.

Wurm, S. A. (1967). 'Linguistics and the prehistory of the south-western Pacific', *The Journal of Pacific History* 2(1): 25–38.

Yang, C. D. (2000). 'Internal and external forces in language change', *Language Variation and Change* 12(3): 231–250.

Yu, G., Lam, T. T.-Y., Zhu, H., and Guan, Y. (2018). 'Two methods for mapping and visualizing associated data on phylogeny using Ggtree', *Molecular Biology and Evolution* 35: 3041–3043.

Zhou, K. and Bowern, C. (2015). 'Quantifying uncertainty in the phylogenetics of Australian numeral systems', *Proceedings of the Royal Society B: Biological Sciences* 282: 20151278.

9780192886491_Book.indb 421 27-Mar-25 23:04:35









CHAPTER 22

INFORMATION-THEORETIC AND MACHINE LEARNING METHODS FOR SEMANTIC CATEGORIZATION

FRANCIS MOLLICA AND NOGA ZASLAVSKY

22.1 Introduction: What makes a useful system of semantic categories?

We live in a remarkably structured world, yet most of this structure goes unremarked. As an illustration, consider a deck of fifty-two playing cards. There are over 4.5×10^{15} possible ways that we can group the cards, yet only a handful of groupings are privileged enough to be named (e.g. Spades or Aces). When we use the cards to play a game like five card poker, there are 52^5 possible five-card hands that could be played; however, only a handful of them have been semantically organized into named categories (e.g. flush, royal flush, straight). Further, while distinctions relevant for poker might also be relevant for rummy, they do not appear to be relevant or worth categorizing at all if you only play solitaire and go-fish. Thus, at the heart of any semantic categorization, there is a requirement to define what makes a *useful* system of categories?

The obvious answer to the question is that we make categories to help us achieve our goals. A reasonable goal for categorization is to predict unobserved properties of objects in our environments (Anderson 1991). For example, we can observe lots of bananas: some yellow, some green, some bitter, and some sweet. Based on these observations, we learn that there is a co-modality between the colour of a banana and its sweetness. Due to this co-modality (Feldman 2012), we can make a category that contains the sweet, yellow bananas so that if we observe a yellow banana we can make a fair prediction that it will be sweet. We can lexicalize this category and call it 'ripe.' When we lexicalize a category, it's called a *semantic category*.

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

While prediction of unobserved features is a reasonable goal, it is not the only goal that people have when building semantic categories. If we only built and lexicalized categories for prediction, we would make a lot of bad predictions as there are a lot of spurious correlations in the environment that we could pick up on (Murphy and Medin 1985). For example, not all birds can fly or defeat Australia in open war. Further, there are an infinite, and thus unstorable, number of ways we could segment the world that might have high predictive ability. Therefore, we want easily acquired categories that balance predictive ability and memory strain in addition to prediction (Bruner et al. 1956).

Along these lines, Rosch (1978) famously argued for two general principles for the formulation of categories: *cognitive economy* and *perceived world structure*. Cognitive economy defines a trade-off between the complexity of a categorization system and the utility of the system, resonating with Zipf's (1949) least-effort principle. A system should not be more complex than it needs to be to achieve its goals. For example, if my goal is to judge and communicate whether a banana is ready to eat, I only need semantic categories for 'ripe' and 'unripe'. On the other hand, if my goals also include baking, I might want a semantic category for 'overripe'. *Perceived world structure* reflects the fact that how we perceive the world and its co-modalities is not a property of some platonic world but a property of cognitive agents. Communities and individuals can have different goals they are trying to achieve in different environments which privilege different segmentations of the world. For example, in a baking society, baking goals might warrant the additional complexity of separating ripe from overripe fruit; whereas, in a non-baking society we might not expect degrees of ripeness to be lexicalized.

If it is true that a society's goals are reflected in their semantic categorization, then studying semantic typology—i.e. the diversity of semantic categories across cultures—provides us leverage to understand the diversity of goals, the diversity of categorization strategies devised to achieve those goals and how goals and semantic categories have co-evolved.

22.1.1 Efficiency of semantic categorization

Recently, it has been argued that a plausible goal for semantic categorization, and language more broadly, is to support efficient communication, and there is a substantial body of evidence supporting this view (for reviews, see Kemp et al. 2018; Gibson et al. 2019). When cast in terms of communication, Rosch's (1978) notion of cognitive economy dovetails nicely with a rich linguistic tradition of using competing motivations to understand language (including von der Gabelentz 1901; Du Bois 1985; Haiman 2010). For example, in Zipf's (1949) formulation, languages trade off between the Force of Unification—i.e. a speaker's desire for a simple system (e.g. having few words)—and the Force of Diversification—i.e. a listener's desire to accurately reconstruct intentions (e.g. by having unique words for every possible world state). Languages can differentially trade off these two pressures to settle on simple, informative yet diverse systems.

9780192886491_Book.indb 424 27-Mar-25 23:04:36

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

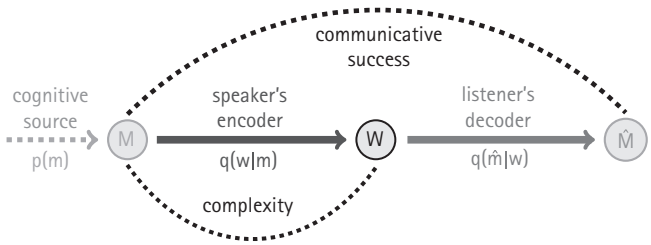


FIGURE 22.1 Application of Shannon's communication problem to semantic categorization, as proposed by Zaslavsky et al. (2018).

Communication model. Speakers select an intended meaning m to be encoded into a word w via a probabilistic encoder q(w|m). The word passes through a (noiseless) communication channel and reaches a listener, who then reconstructs the intended meaning $\hat{m}$. Communicative success is inversely related to the error between the speaker's intended meanings and listener's reconstructions. The complexity of the system is roughly the number of bits that are required for communication.

In the past decade, we have seen several approaches to formalizing this communicative efficiency trade-off and evaluating its adequacy at characterizing semantic typology. These efficient communication models have been useful in explaining the constrained variation in the world's lexicons across multiple semantic domains, including kinship (Kemp and Regier 2012), colour (Zaslavsky et al. 2018), number (Xu et al. 2020), quantifiers (Steinert-Threlkeld 2020), personal pronouns (Zaslavsky et al. 2021b), modals (Uegaki et al. 2023), Boolean connectives (Uegaki 2022), grammatical features (Mollica et al. 2021; Denić et al. 2022), and spatial prepositions (Chen et al. 2022). While all of these studies have considered a trade-off between a pressure for *communicative simplicity* and a pressure for *communicative success*, these trade-offs have often been formalized in different ways, leading to different, and possibly contradictory, predictions. One formal theoretical framework that offers a consistent, domain-general formulation of this trade-off is based on the Information Bottleneck (IB) principle (Tishby et al. 1999; Zaslavsky et al. 2018), which is a general principle that is rooted at the intersection of information theory and machine learning (e.g. Chechik et al. 2005; Shamir et al. 2010; Tishby and Zaslavsky 2015). This framework is explicitly grounded in Shannon's (1948) communication problem (illustrated in Figure 22.1)—i.e. how to reliably transmit information from a source (e.g. speaker's mind) to a destination (e.g. listener's mind) using a code (e.g. language)—and it makes quantitative theoretical predictions for the fine-grained structure of semantic categories across languages and their evolution (Zaslavsky 2020).

Here, we review this framework in detail. We begin by comparing several definitions of communicative complexity and success that were proposed in the literature and discuss them in relation to the IB principle. In section 22.2, we review the IB framework in greater detail, including a demonstration in the case of colour and a general discussion of the kind of research questions that this framework may help to address. In section

9780192886491_Book.indb 425 27-Mar-25 23:04:36

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

22.3, we review recent applications of this framework to emergent communication in multi-agent systems.

22.1.1.1 Complexity

There are several working operationalizations of complexity for the purpose of linguistic analysis. One intuitive notion of complexity is the size of the lexicon (Regier et al. 2015). The more words[1] we have, the harder it will be to retrieve them from memory, however, the more precisely we can communicate in the absence of a context. From a methodological standpoint, this formalization is easy to deploy as it does not require a rich understanding of the semantic space, and linguistic forms are relatively easy to elicit. Nevertheless, there are considerable limitations. First, as a matter of face validity, the semantics of the system are completely irrelevant; whereas, there are strong intuitions that some semantic distinctions (e.g. parallel-/cross-cousins) are more complex than others (e.g. maternal/paternal cousins). Second, words are often co-lexified, such that one word is assigned to more than one semantic category when the different contexts of use easily disambiguate (Piantadosi et al. 2012). In these cases, systems with co-lexified words would distort the communicative trade-off as they achieve greater communicative success than systems without co-lexification while failing to account for the complexity of mapping words to multiple semantic categories.

Another approach has been to calculate the Minimum Description Length (MDL) of each semantic category in terms of some semantic metalanguage (e.g. Kemp and Regier 2012; Steinert-Threlkeld 2020). The underlying intuition of MDL approaches is to capture the algorithmic complexity of the meaning (Kolmogorov 1963; Chater and Vitányi 2003; Knowlton 2021). More complex categories will require longer MDLs. For example, Kemp and Regier (2012) motivated an anthropological description language for kinship terms based on the logic of genealogical relationships (Kroeber 1909; Greenberg 1949; Gould 2000). A concept like 'mother' could be described as $\mathit{female}(x) \wedge \mathit{parent}(x, y)$, which is shorter and thus less complex than 'aunt', described as $\exists z\, \mathit{parent}(z, y) \wedge \mathit{sibling}(x, z) \wedge \mathit{female}(x)$. A major boon of this approach is that complexity can be directly grounded in semantic or developmental processes (e.g. Xu et al.'s 2020 number model), providing explainability and the means for strong generalization across tasks. At the same time, this approach requires sophisticated knowledge of semantics that we rarely possess for most of the world's languages. Further, it is virtually impossible (i.e. computationally intractable) to determine the unique minimum description for semantic categories (e.g. see Burling (1964) for kinship non-identifiability).

From an information-theoretic perspective (Bialek et al. 2001; Gilad-Bachrach et al. 2003), the complexity of a communication system is measured by $I(M;W)$, the mutual information between messages (or meanings) and codewords. This measure of complexity has been mathematically derived by Shannon (1959) as a solution to his general problem of communication, and a similar derivation is employed in the IB framework

[1] We ignore the complexity of defining what is meant by a *word* and simply note that this notion applies to any level of linguistic form (e.g. morphemes).

9780192886491_Book.indb 426 27-Mar-25 23:04:37

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

(Tishby et al. 1999; Gilad-Bachrach et al. 2003). Intuitively, it quantifies the minimal number of bits, on average, that are required for communicating with a sufficiently small error. This measure is tightly related to the MDL principle (Bialek et al. 2001), but does not require an *a priori* specification of a semantic metalanguage and can therefore be applied in a domain-general manner. In addition, while this informational term is also related to the lexicon size, such that higher complexities may require larger lexica, it is a softer measure of complexity that is sensitive also to the probabilistic patterns of language use. For example, a categorization system with ten uniformly used words (or categories) would be more complex than a system with the same number of words in which one word is used 50 per cent of the time. This property has been shown to be crucial for predicting empirical phenomena such as soft category boundaries and inconsistent naming across speakers as well as reconciling continuous and discrete aspects of language evolution (Zaslavsky et al. 2018, 2022).

22.1.1.2 Communicative success

What does it mean to succeed at communication? As should be readily apparent to everyone, success is a subjective notion. At face value, communicative success is achieved when the listener recovers the intended meaning of the speaker. If the speaker and listener are willing to tolerate some degree of error, which is typically the case, then defining communicative success requires a measure of similarity, or distortion, between the speaker's and listener's meaning representations. Early models of communicative efficiency (e.g. Kemp and Regier 2012; Regier et al. 2015) made a simplifying assumption that the speaker has a precise representation of the intended referent. In those cases, a natural formulation of communicative success is taken to be the expected log likelihood that the listener assigns to the speaker's intended meaning. This approach captures the gradience of communicative success in terms of the listener's uncertainty over referents. A smaller listener log likelihood reflects greater uncertainty about the speaker's intention. This formulation is also closely related to communicative utility in rational models of recursive pragmatic reasoning (Frank and Goodman 2012; and see also Zaslavsky et al. 2021a). In contrast to the expected log likelihood, some models (e.g. Steinert-Threlkeld 2020) operationalize communicative success as a binary utility function, where the probability of success is given by such a model of pragmatic communication.

In reality, the speaker certainty assumption is too restrictive as speaker's intentions are not always precise, and thus provide an additional source of noise. Relaxing this assumption naturally leads to the Kullback-Leibler (KL) divergence—which is the expected log likelihood ratio between the speaker's and listener's beliefs about the target referent—as the measure of distortion, or communicative cost. Communicative success is then negatively related to this KL divergence (Zaslavsky et al. 2018), and trading it off with the informational complexity measure discussed earlier leads to an optimization problem which has been known in the machine learning and information theory literature as the Information Bottleneck (IB) principle (Tishby et al. 1999).

9780192886491_Book.indb 427 27-Mar-25 23:04:37

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Clearly, transmitting information from a speaker to a listener is required for communicative success; however, the real goal of language use is often not to convey a particular message to a listener but to enact change in the world. This notion of successful communication is not captured by simple likelihood-based measures. Tucker et al. (2022b) therefore suggest an integrative framework that trades off communicative complexity and informativeness, as in the IB principle, as well as task-specific utility. It is worth noting, however, that even without explicitly taking into account task-specific utilities, these general communicative pressures towards minimizing communicative complexity and maximizing informativeness have been shown empirically to provide compelling explanations for the semantic typology of several domains (Zaslavsky et al. 2018, 2019, 2021b; Mollica et al. 2021).

22.1.1.3 Communicative need

Before elaborating on the application of the IB principle to semantic categorization, we wish to note that most of the aforementioned measures of communicative complexity and success, including in IB, require a specification of a need distribution (Kemp et al. 2018; Zaslavsky et al. 2020). That is, the probability with which each element in an environment is communicated. Incorporating this component into the efficiency formulation increases the model's ecological validity, such that the success of more needed speaker intentions is privileged compared to less needed intentions (Gibson et al. 2017). However, estimating communicative need is a challenge. Several methods for addressing this challenge have been proposed in the literature, including corpus-based methods (Kemp and Regier 2012; Zaslavsky et al. 2020), maximum-entropy based methods (Zaslavsky et al. 2020; Twomey et al. 2021), extracting statistics from naturalistic images (Gibson et al. 2017), and experimental elicitation of human judgements (Gao and Regier 2022). Each one of these methods provides a way to estimate need probabilities given a set of assumptions about the underlying factors that may shape communicative need and the type of available data that may inform the estimation process. While none of these methods can guarantee an accurate estimation of the actual need probabilities employed by a population, in practice, they have usefully informed the instantiation of many efficient communication models.

An important modelling decision in this context is whether to assume a universal need distribution or a language-specific need distribution. On the one hand, a universal need distribution is clearly a simplifying assumption, as communicative needs are likely to vary across cultures. On the other hand, this simplifying assumption has been shown to be very powerful, leading to simpler models that can account remarkably well for cross-language semantic variation (Zaslavsky et al. 2020, 2021b; Mollica et al. 2021). This suggests that at least some aspects of communicative need may be shared across cultures (Zaslavsky et al. 2021b), although incorporating language-specific needs may indeed be useful in some cases (Twomey et al. 2021; Gao and Regier 2022) and warrant additional model complexity.

9780192886491_Book.indb 428 27-Mar-25 23:04:38

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

22.2 The Information Bottleneck (IB) framework for semantic categories

While we are primarily concerned with linguistic representations in this chapter, many cognitive representations can be viewed as capacity limited communication, or Rate-Distortion problems (Berger 1971; Bialek et al. 2006; Sims 2016). As cognitive agents, we want to achieve goals in complex environments. For any goal, a subset of the environmental information is useful for goal attainment. Therefore, we want to cognitively represent as much of the useful information from the environment as possible. At the same time, we often do not have the memory resources to store all of the useful information. Further, we may even want to strategically abstract over some of the specific information to ensure robust generalization (O'Connor 2014; Gershman 2017; Sims 2018). Regardless of the reason, storing less goal-relevant information in a context will hinder our goal attainment in that context. As a result, we are faced with a trade-off between simple representations that are within memory capacity, generalize robustly yet hinder goal attainment, and informative representations that require complicated representations but support goal attainment. Rate-distortion theory (RDT; Shannon 1959; Berger 1971), the branch of information that addresses the problem of obtaining the most concise representations given a fidelity criteria, offers a natural mathematical framework for formulating and analysing such trade-offs. However, RDT does not address the problem of specifying a fidelity criteria for communicative goals. The IB principle emerges as a special case of RDT when the fidelity criteria, or distortion, is taken to be the KL divergence between the sender's and receiver's belief states (Harremoës and Tishby 2007; Zaslavsky 2020), as discussed in section 22.1.1.2. Therefore, the IB principle lends a theoretically rich mathematical framework for modelling and analysing semantic systems across languages.

To illustrate the IB framework for semantic categorization, let us consider its application to the task of colour communication (Zaslavsky et al. 2018). Instantiating the communication problem in Figure 22.2a requires (1) a representation of the source domain which defines the set of speaker's mental representations, or meanings, m (2) the communicative need for each meaning $p(m)$, and (3) the set of possible words w.

Taking each in turn, the source domain in this example is colour, the *perception* of electromagnetic radiation. Specifying the underlying representation of the domain requires not only the specification of a set of elements in the environment that can be communicated, but also embedding them in a similarity space in order to measure the degree of communicative success. In many domains there are no pre-established similarity spaces, in which case approximation methods are required, such as grounding the space in non-linguistic human similarity judgements (Xu et al. 2016) or in a set of perceptual or conceptual features (Eisape et al. 2020; Zaslavsky et al. 2021b). In the case of colour, however, it is possible to use a well-established perceptual colour space

9780192886491_Book.indb 429 27-Mar-25 23:04:38

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

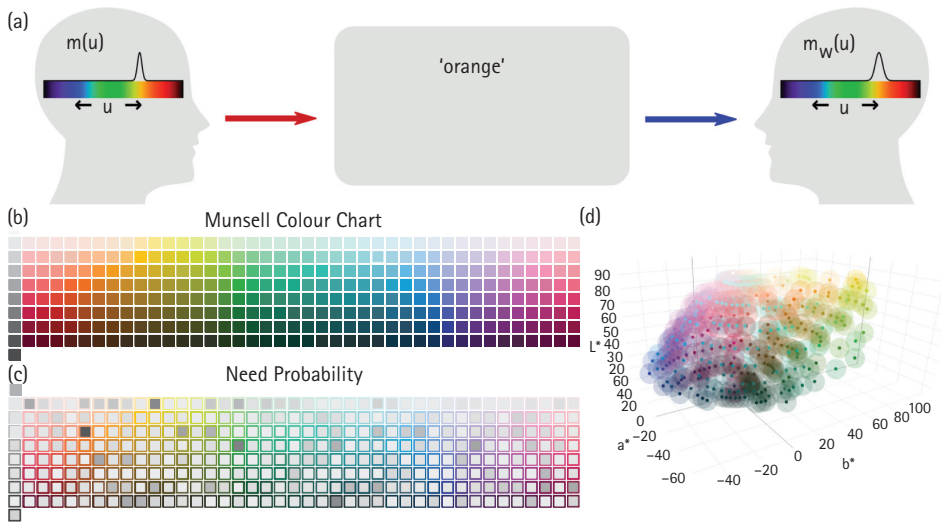


FIGURE 22.2 a) Instantiation of the communication model in the domain of colour, where each colour is mentally represented by a probability distribution over a perceptual colour space; b) the chips in the Munsell colour chart; c) the communicative need *p(m)* is reflected in the shading of each colour chip; d) CIElab colour space resembling speaker intentions *m(u)* for each of the colour chips in the Munsell colour chart.

called CIELAB (Regier et al. 2007), which is given by three perceptual dimensions (Figure 22.2d): lightness (L^*), and hue and saturation which are encoded in polar coordinates (a^*,b^*). CIELAB was designed for perceptual uniformity, such that a small equal movement along any dimension, with respect to the Euclidean distance, results in a just noticeable difference. While the CIELAB space is continuous, semantic typologists cannot use the infinite space when doing fieldwork. Instead they use a set of paint chips, Munsell colour chips (see Figure 22.2b), which have been carried into the field as part of the World Color Survey (WCS; Berlin and Kay 1969, Cook et al. 2005). These chips when represented in CIELAB coordinates are a decent, although non-uniform, sampling of the perceptual colour space and have therefore been used in the literature (e.g. Regier et al. 2015; Zaslavsky et al. 2018) as a stand-in for all the possible colour intentions a speaker would like to convey. Further, a speaker can be uncertain or unable to distinguish coordinates in the perceptual colour space due to perceptual noise (see O'Connor 2014 for the advantages of perceptual uncertainty). Therefore, Zaslavsky et al. (2018) define the speaker's intended meaning by a probability distribution $m(u)$ over colour space,[2] which, in this case, takes the form of a normal distribution centred on the coordinates for the target colour chip u_t, with a standard deviation reflecting the just noticeable difference in CIELAB.

[2] Note that perceptual certainty is a special case of this framework as these distributions can be point distributions.

9780192886491_Book.indb 430 27-Mar-25 23:04:40

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Next, to specify the communicative need distribution, $p(m)$, Zaslavsky et al. (2018) employed a minimal-assumptions estimation method based on the notion of least-informative priors.[3] Intuitively, this method aims to find the prior $p(m)$ that maintains the least amount of information about the posterior $p(w|m)$, corresponding to a naming distribution. To ensure that this method does not overfit to any specific language, they considered a universal need distribution by averaging language-specific least-informative need distributions. Furthermore, they evaluated their model using k-fold cross-validation, such that each language was evaluated using a need distribution that was inferred from a subset of randomly selected languages excluding that language. Their resulting average least-informative need distribution is shown in Figure 22.2c. It should be noted that while the IB framework does not require the universal need-distribution assumption, it is a common assumption that has been shown to be empirically powerful, and can be relaxed when desired (e.g. Zaslavsky et al. 2021b; Twomey et al. 2021).

Finally, the vanilla IB framework does not explicitly take into account the specific form of each word w, but rather treats words as arbitrary discrete symbols. Therefore, defining the set of possible words amounts to specifying the lexicon size. In practice, Zaslavsky (2020) showed that it is not necessary to impose a hard constraint on the lexicon size, as structural phase transitions in the lexicon size naturally emerge in the IB framework. More recent studies have extended this framework to explicitly account for linguistic forms (Mollica et al. 2021; Tucker et al. 2022a), however in domains other than colour.

With this instantiation of Shannon's communication model for colour it is possible to compute the IB trade-off, as discussed in the previous section. Recall that complexity in IB is defined as the mutual information between speaker's word w and intended meaning m, formally denoted by $I(M;W)$, and communicative success (or accuracy) is defined as the expected KL divergence between the speaker's intended meaning m and the listener's reconstructed meaning $\widehat{m}$, formally denoted by $\mathbb{E}[D[M||\widehat{M}]]$. The IB trade-off between complexity and communicative success is governed by a trade-off parameter β, which yields the following optimization problem:

$$\text{minimize} \quad I\left(W;M\right)+\beta\mathbb{E}\left[D[M||\widehat{M}]\right], \tag{1}$$

where the optimization is taken over all feasible speaker encoders $q(w|m)$ and listener's decoders $q(\widehat{m}|w)$ (see Figure 22.1). Given a trade-off β, this optimization problem can be solved using the IB method (Tishby et al. 1999), which is a variant of an optimization method known as the Blahut-Arimoto algorithm (Blahut 1972; Arimoto 1972), in combination with a method called deterministic annealing (Rose 1998), which tracks the solution by gradually increasing (or decreasing) the value of β. This yields a continuous evolutionary process for the optimal IB communication systems (Zaslavsky 2020) that

[3] See also Zaslavsky et al. (2020) for similar results based on a maximum-entropy estimation method, as well as a systematic evaluation of several estimation methods.

9780192886491_Book.indb 431 27-Mar-25 23:04:44

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

spans the IB Pareto frontier, also known as the IB curve (Gilad-Bachrach et al. 2003), which, in this context, provides the theoretical limit of efficiency for semantic systems (see Figure 22.3, black curve).

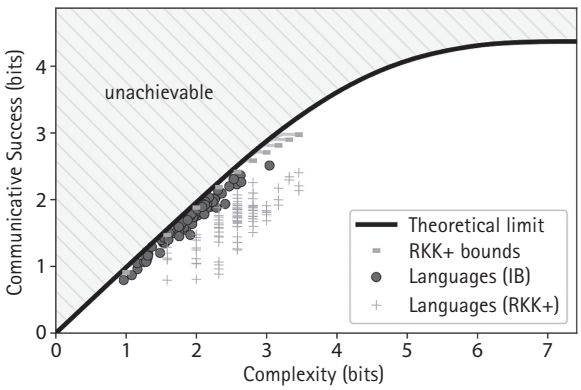


FIGURE 22.3 Complexity–accuracy values for all languages in the World Color Survey (Cook et al. 2005), as well as English (Lindsey and Brown 2014), according to IB (circular dots) and Regier et al.'s (2015) formalization of the trade-off (crosses).

The IB curve (black) defines the theoretical limit of achievable trade-offs. Natural languages under the IB near optimally trade-off success and complexity, i.e. closer to the black line.

Source: Zaslavsky et al. (2018).

With this normative model for rational analysis (Anderson 1989), several questions about semantic categories can be entertained:

1. **Are the semantic categories of attested languages (near) optimal for communication?** Zaslavsky et al. (2018) derived two quantitative theoretical predictions to address this question. First, if languages are near-optimal, then their deviation from optimality—intuitively, their distance from the IB theoretical limit—should be small. This can be evaluated by scoring each attested language on complexity and communicative success (Figure 22.3, blue dots) and comparing them with the optimal IB trade-offs (Figure 22.3, black curve). As shown in Figure 22.3, attested languages lie very close to the theoretical limit, which means that they are near-optimal in the IB sense. While this measure captures nearness in terms of the trade-off properties, it does not capture nearness in terms of the topology and/or behaviour of the underlying language. Therefore, Zaslavsky et al. (2018) formulated a second theoretical prediction that captures the behavioural similarity between the attested and optimal systems. To this end, they proposed a measure called the generalized normalized information distance (gNID), which, intuitively, measures how much information about speaker intentions is not shared across words in two different languages. For example, *blue* in English and *azul* in Spanish behave similarly, contributing to a low gNID; whereas, *blue* in English and синий (dark blue) in Russian behave differently, resulting in greater gNID. Using this measure,

9780192886491_Book.indb 432 27-Mar-25 23:04:44

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Zaslavsky et al. (2018) were able to quantify the apparent similarity between the optimal and attested colour naming systems and show that the optimal systems are human-like in the sense that they are similar (achieve low gNID) to attested systems. In addition, they showed that attested systems are locally optimal with respect to a set of hypothetical variants that preserve the same category structure as the attested languages but are rotated along the hue dimension. Importantly, the IB framework is not specific to colour, and indeed, subsequent work has showed converging evidence in support of the IB-optimality predictions in a variety of semantic domains (e.g. Zaslavsky et al. 2019, 2021b; Mollica et al. 2021; Chen et al. 2022).

2. **Is it possible to discriminate between attested and hypothetical languages based on their efficiency?** To answer this question, we need to generate a large and representative set of possible but unattested systems so that we can evaluate if deviation from optimality is a good predictor of whether a system is attested or not. This method can also predict which kinds of system we would expect to see if we had a more extensive survey of the world's languages. For example, Mollica et al. (2021) adopted this approach to show that the attested grammatical morpheme systems for tense, nominal number, and evidentiality can be reliably discriminated from unattested but logically possible deterministic or permuted marking systems based on their efficiency (or near-optimality) scores alone.
3. **Assuming attested languages are communicatively efficient, which sets of underlying cognitive assumptions are better explanations of typological diversity?** As in rational analysis (Anderson 1989), we often agree on the computational-level goal (Marr 1982)—in this case, efficient communication; however, researchers may disagree on the constraints that influence how the goal is attained. For example, Regier et al. (2015, RKK) formalized the colour communication problem using number of terms as a complexity measure, in contrast to the IB informational complexity measure. As a result, the RKK formulation predicts that word–meaning mappings are deterministic, even without making such an assumption explicit in the communication model. In the case of colour, it is straightforward to tease these two efficiency principles apart as the data clearly reflects probabilistic mappings, including soft category structure and inconsistent naming patterns across speakers, thus lending further empirical support to the IB principle. This difference between the two formulations is also apparent quantitatively, when comparing attested languages with the optimal systems that each formulation predicts (Figure 22.3). Another type of assumption that can be tested are related to the need distribution and underlying representation of the domain. For example, Zaslavsky et al. (2020) used the IB framework to systematically evaluate different types of estimation methods for communicative need. Interestingly, they found that methods that are based on linguistic use outperform methods that are based on natural image statistics, at least in the domain of colour, suggesting that language use is a better predictor of communicative need compared to the statistics of colours in the environment. Eisape et al. (2020) used

9780192886491_Book.indb 433 27-Mar-25 23:04:44

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

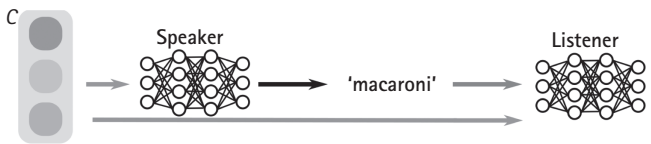


FIGURE 22.4 An example of a reference game played by a multi-agent system.

One agent is assigned a speaker role and tasked with communicating a target referent in a given environment *C* to a listener agent. The speaker encodes the target referent into a word (e.g. *macaroni*), which is then transmitted to the listener agent who needs to solve some downstream task, such as selecting the target referent.

a similar approach in order to test which computer vision models may offer useful representations for understanding the semantic categorization of visual objects, and Zaslavsky et al. (2021b) extended this approach to testing hypotheses about underlying mental conceptual features in the domain of personal pronouns.

22.3 The emergence of efficient categories in multi-agent systems

While communicative efficiency models have achieved considerable success, one criticism of this body of work is the lack of an explanatory process or mechanism for language evolution (Levinson 2012). While strictly true, models of language evolution often require a bias for information loss (simplicity) and a bias for informativity (Spike et al. 2017). The communicative efficiency trade-off potentially provides an overarching explanation for why these biases are present in every model of language evolution and, further, provides a reason for the constrained variation observed cross-linguistically. In a popular line of more mechanistic models (e.g. Carr et al. 2020), the biases are often explained in terms of cultural evolution—i.e. the iterated acquisition and social transmission of a cognitive technology (Heyes 2018).

While the existing dynamics of cultural evolution in humans is a source of ongoing debate[4] there is a rich literature on the emergence of semantic categories in multi-agent systems playing reference games (Steels and Belpaeme 2005; Dowman 2007; Loreto et al. 2012). More recently, emergent communication has been explored using contemporary deep learning methods (e.g. Foerster et al. 2016; Lazaridou et al. 2017; Mordatch and Abbeel 2018).

In a multi-agent system, there is an environment of stimuli (e.g. colour chips) and a set of agents. At each time step, agents are randomly assigned to the roles of speaker

[4] Some researchers have argued that simplicity is an acquisition bias and informativity is a communication bias (e.g. Kirby et al. 2015); whereas, others have argued that, in some circumstances, acquisition is sufficient (Gyevnar et al. 2022).

9780192886491_Book.indb 434 27-Mar-25 23:04:45

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

and listener. Speakers are given a colour chip in a particular context and asked to produce a word to refer to a target chip in the context. Listeners are given the word and the context and asked to select the target chip in the context. Over repeated games, the population of agents converges on a linguistic system. It has been illustrated that these models recapitulate presumed historical patterns in semantic typology (e.g. Dowman 2007; Baronchelli et al. 2010; Loreto et al. 2012).

Recently, multi-agent systems have been connected to efficient communication. Kågebäck et al. (2020) and Chaabouni et al. (2021) showed that simple multi-agent reinforcement learning dynamics can converge to efficient category systems in the domain of colour. That said, reinforcement learning dynamics is a deviation from IB dynamics as the goal of the representation is not communication (i.e. reconstruction of the speaker distribution) but rather performance in a downstream task, such as selecting a target referent (see Figure 22.4). To the extent that the agents' environment is realistic, this is a boon because our linguistic systems should encode relevant contextual information about our goals. At the same time, it is easy to overfit to the training contexts, in which case we do not learn human-like systems that adaptively generalize well out of context. Importantly, these reinforcement learning dynamics can only account for cross-linguistic variation by appealing to environmental factors, including noise in communication or differential costs for communicative errors.

In response to these criticisms, Tucker et al. (2022b) proposed a new approach that integrates the IB framework with utility maximization reinforcement learning approaches, and suggested a new method, Vector Quantified-Variational Information Bottleneck (VQ-VIB), for implementing and training agents in a more cognitively motivated way. This integrative framework allows agents to continuously evolve over time via annealing, as in Zaslavsky et al. (2018, 2022), but at the same time, it also takes into account task-specific utilities. Tucker et al. (2022b) evaluated this approach in two cooperative navigation tasks and a colour reference game, and Tucker et al. (2022a) extended this line of work to open-domain communication across a wide range of visual inputs. They show that this integrative framework has several key advantages: the emergent VQ-VIB communication systems efficiently approximate the IB theoretical bound; VQ-VIB agents are adaptive and evolve useful grounded representations of the environment; and this method outperforms previous methods by showing a better fit to actual human languages, better generalization to out-of-domain inputs, and better convergence rates.

22.4 Summary

To the extent that semantic categories are useful, they reflect the goals of a culture. Using cognitive representations to achieve cognitive goals entails an efficiency

9780192886491_Book.indb 435 27-Mar-25 23:04:45

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

trade-off between the complexity of the representations and the success of goal attainment. Communication is an obvious goal for semantic categories and the communicative efficiency trade-off is well studied with several formalizations. Of recent formalizations, the IB formalization is the theoretical bound for the communicative goal and can be used to address a variety of questions including 1) whether semantic categories are (near) optimal for communication, 2) if we can discriminate between attested and unattested languages based on the principles of efficient communication, and 3) which of two sets of assumptions provide a better explanation of typological diversity. While current research on the IB framework for semantic categorization has focused on incorporating other goals (e.g. contextualized reference) in multi-agent systems, moving towards the future, we will need richer datasets (e.g. for diachronic evaluation) and more work on how the multiple goals that we want to attain as cognitive agents constrain each other as we build, transmit, and use cognitive technologies.

References

Anderson, J. R. (1989). 'A rational analysis of human memory', in H. L. Roediger III and F. Craik (eds.), *Varieties of Memory and Consciousness: Essays in Honour of Endel Tulving*, 195. New York: Psychology Press.

Anderson, J. R. (1991). 'The adaptive nature of human categorization', *Psychological Review* 980(3): 409.

Arimoto, S. (1972). 'An algorithm for computing the capacity of arbitrary discrete memoryless channels', *IEEE Transactions on Information Theory* 180(1): 14–20.

Baronchelli, A., Gong, T., Puglisi, A., and Loreto, V. (2010). 'Modeling the emergence of universality in color naming patterns', *Proceedings of the National Academy of Sciences* 1070(6): 2403–2407.

Berger, T. (1971). *Rate Distortion Theory; A Mathematical Basis for Data Compression*. Englewood Cliffs, NJ: Prentice-Hall.

Berlin, B. and Kay, P. (1969). *Basic Color Terms: Their Universality and Evolution*. Berkeley, Los Angeles: University of California Press.

Bialek, W., Nemenman, I., and Tishby, N. (2001). 'Predictability, complexity, and learning', *Neural Computing* 130(11): 2409–2463, ISSN 0899-7667. doi: 10.1162/089976601753195969.

Bialek, W., de Ruyter van Steveninck, R. R., and Tishby, N. (2006). 'Efficient representation as a design principle for neural coding and computation', *IEEE International Symposium on Information Theory* 659–663.

Blahut, R. (1972). 'Computation of channel capacity and rate-distortion functions', *IEEE Transactions on Information Theory* 180(4): 460–473.

Bruner, J. S., Goodnow, J., and Austin, G. A. (1956). 'The process of concept attainment', *A Study of Thinking*. New York: John Wiley and Sons.

Burling, R. (1964). 'Cognition and componential analysis: God's truth or hocus pocus?', *American Anthropologist* 660(1): 20–28.

Carr, J. W., Smith, K., Culbertson, J., and Kirby, S. (2020). 'Simplicity and informativeness in semantic category systems', *Cognition* 202: 104289.

9780192886491_Book.indb 436 27-Mar-25 23:04:45

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Chaabouni, R., Kharitonov, E., Dupoux, E., and Baroni, M. (2021). 'Communicating artificial neural networks develop efficient color-naming systems', *Proceedings of the National Academy of Sciences* 1180(12): e2016569118, doi: 10.1073/pnas.2016569118.

Chater, N. and Vitányi, P. (2003). 'Simplicity: A unifying principle in cognitive science?', *Trends in Cognitive Sciences* 70(1): 19–22.

Chechik, G., Globerson, A., Tishby, N., and Weiss, Y. (2005). 'Information bottleneck for Gaussian variables', *Journal of Machine Learning Research* 60(6): 165–188, http://jmlr.org/papers/v6/chechik05a.html.

Chen, S., Futrell, R., and Mahowald, K. (2022). 'Investigating information-theoretic properties of the typology of spatial demonstratives', in *Proceedings of the 4th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, 94–95.

Cook, R. S., Kay, P., and Regier, T. (2005). 'The world color survey database', in H. Cohen and C. Lefebvre (eds.), *Handbook of Categorization in Cognitive Science*, 223–241. Amsterdam: Elsevier.

Denić, M., Steinert-Threlkeld, S., and Szymanik, J. (2022). 'Indefinite pronouns optimize the simplicity/informativeness trade-off', *Cognitive Science* 460(5): e13142.

Dowman, M. (2007). 'Explaining color term typology with an evolutionary model', *Cognitive Science* 310(1): 99–132.

Du Bois, J. W. (1985). 'Competing motivations', in J. Haiman (ed.), *Iconicity in Syntax*, 343–365. Amsterdam: Benjamins.

Eisape, T., Levy, R., Tenenbaum, J. B., and Zaslavsky, N. (2020). 'Toward human-like object naming in artificial neural systems', in *Bridging Artificial Intelligence and Cognitive Science workshop*. Addis Ababa: ICLR.

Feldman, J. (2012). 'Symbolic representation of probabilistic worlds', *Cognition* 1230(1): 61–83.

Foerster, J. N., Assael, Y. M., de Freitas, N., and Whiteson, S. (2016). 'Learning to communicate with deep multi-agent reinforcement learning', in *Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16,* 2145–2153, Red Hook, NY: Curran Associates Inc.

Frank, M. C. and Goodman, N. D. (2012). 'Predicting pragmatic reasoning in language games', *Science* 3360(6084): 998–998.

Gao, S. and Regier, T. (2022). 'Culture, communicative need, and the efficiency of semantic categories', in *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Gershman, S. J. (2017). 'On the blessing of abstraction', *The Quarterly Journal of Experimental Psychology* 700(3): 361–365.

Gibson, E., Futrell, R., Jara-Ettinger, J., et al. (2017). 'Color naming across languages reflects color use', *Proceedings of the National Academy of Sciences* 1140(40): 10785–10790.

Gibson, E., Futrell, R., Piantadosi, S. P., et al. (2019). 'How efficiency shapes human language', *Trends in Cognitive Sciences* 230(5): 389–407.

Gilad-Bachrach, R., Navot, A., and Tishby, N. (2003). 'An information theoretic tradeoff between complexity and accuracy', in B. Schölkopf and M. K. Warmuth (eds.), *Learning Theory and Kernel Machines,* 595–609. Berlin, Heidelberg: Springer. ISBN 978-3-540-45167-9.

Gould, S. H. (2000). *A New System for the Formal Analysis of Kinship*. Lanham, MD: University Press of Amer.

Greenberg, J. H. (1949). 'The logical analysis of kinship', *Philosophy of Science* 160(1): 58–64.

Gyevnar, B., Dagan, G., Haley, C., et al. (2022). 'Communicative efficiency or iconic learning: Do acquisition and communicative pressures interact to shape colour-naming systems?', *Entropy* 240(11): 1542.

9780192886491_Book.indb 437 27-Mar-25 23:04:45

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Haiman, J. (2010). 'Competing motivations', in J. J. Song (ed.), *The Oxford Handbook of Linguistic Typology*, 148–165. Oxford: Oxford University Press.

Harremoës, P. and Tishby, N. (2007). 'The information bottleneck revisited or how to choose a good distortion measure', 7: 566–570, doi: 10.1109/ISIT.2007.4557285.

Heyes, C. (2018). *Cognitive Gadgets: The cultural Evolution of Thinking*. Boston, MA: Harvard University Press.

Kågebäck, M., Carlsson, E., Dubhashi, D., and Sayeed, A. (2020). 'A reinforcement-learning approach to efficient communication', *PLOS ONE* 150(7): 1–26.

Kemp, C. and Regier, T. (2012). 'Kinship categories across languages reflect general communicative principles', *Science* 3360(6084): 1049–1054.

Kemp, C., Xu, Y., and Regier, T. (2018). 'Semantic typology and efficient communication', *Annual Review of Linguistics* 4: 109–128, doi: 10.1146/annurev-linguistics-011817-045406.

Kirby, S., Tamariz, M., Cornish, H., and Smith, K. (2015). 'Compression and communication in the cultural evolution of linguistic structure', *Cognition* 141: 87–102.

Knowlton, T., Hunter, T., Odic, D., et al. (2021). 'Linguistic meanings as cognitive instructions', *Annals of the New York Academy of Sciences* 15000(1): 134–144.

Kolmogorov, A. N. (1963). 'On tables of random numbers', *Sankhyā: The Indian Journal of Statistics, Series A* 25(4): 369–376.

Kroeber, A. L. (1909). 'Classificatory systems of relationship', *The Journal of the Royal Anthropological Institute of Great Britain and Ireland* 39: 77–84.

Lazaridou, A., Peysakhovich, A., and Baroni, M. (2017). 'Multi-agent cooperation and the emergence of (natural) language', in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net*.

Levinson, S. C. (2012). 'Kinship and human thought', *Science* 3360(6084): 988–989.

Lindsey, D. T. and Brown, A. M. (2014). 'The color lexicon of American English', *Journal of Vision* 140(2): 17.

Loreto, V., Mukherjee, A., and Tria, F. (2012). 'On the origin of the hierarchy of color names', *Proceedings of the National Academy of Sciences* 1090(18): 6819–6824.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Cambridge, MA: MIT Press.

Mollica, F., Bacon, G., Zaslavsky, N., et al. (2021). 'The forms and meanings of grammatical markers support efficient communication', *Proceedings of the National Academy of Sciences* 1180(49): e2025993118, doi: 10.1073/pnas.2025993118.

Mordatch, I. and Abbeel, P. (2018). 'Emergence of grounded compositional language in multi-agent populations', in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'18/IAAI'18/EAAI'18*. AAAI Press.

Murphy, G. L. and Medin, D. L. (1985). 'The role of theories in conceptual coherence', *Psychological Review* 920(3): 289.

O'Connor, C. (2014). 'The evolution of vagueness', *Erkenntnis* 79: 707–727.

Piantadosi, S. T., Tily, H., and Gibson, E. (2012). 'The communicative function of ambiguity in language', *Cognition* 1220(3): 280–291.

Regier, T., Kay, P., and Khetarpal, N. (2007). 'Color naming reflects optimal partitions of color space', *Proceedings of the National Academy of Sciences* 1040(4): 1436–1441, doi: 10.1073/pnas.0610341104.

9780192886491_Book.indb 438 27-Mar-25 23:04:45

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Regier, T., Kemp, C., and Kay, P. (2015). 'Word meanings across languages support efficient communication', in B. MacWhinney and W. O'Grady (eds.), *The Handbook of Language Emergence*, 237–263. New York: Wiley-Blackwell.

Rosch, E. (1978). *Cognition and Categorization, Chapter Principles of categorization*. Mahwah, NJ: Lawrence Erlbaum.

Rose, K. (1998). 'Deterministic annealing for clustering, compression, classification, regression, and related optimization problems', *Proc. IEEE* 860(11): 2210–2239, doi: 10.1109/5.726788.

Shamir, O., Sabato, S., and Tishby, N. (2010). 'Learning and generalization with the Information Bottleneck', *Theoretical Computer Science* 4110(29-30): 2696–2711.

Shannon, C. E. (1948). 'A mathematical theory of communication', *Bell System Technical Journal* 27: 379–423.

Shannon, C. E. (1959). 'Coding theorems for a discrete source with a fidelity criterion', *IRE Nat. Conv. Rec* 40(142–163): 1.

Sims, C. R. (2016). 'Rate–distortion theory and human perception', *Cognition* 152: 181–198.

Sims, C. R. (2018). 'Efficient coding explains the universal law of generalization in human perception', *Science* 3600(6389): 652–656.

Spike, M., Stadler, K., Kirby, S., and Smith, K. (2017). 'Minimal requirements for the emergence of learned signaling', *Cognitive Science* 410(3): 623–658.

Steels, L. and Belpaeme, T. (2005). 'Coordinating perceptually grounded categories through language: A case study for colour', *Behavioral and Brain Sciences* 280(4): 469–489.

Steinert-Threlkeld, S. (2020). 'Quantifiers in natural language optimize the simplicity/informativeness trade-off', in *Proceedings of the 22nd Amsterdam colloquium.* Institute for Logic, Language and Computation at the University of Amsterdam.

Tishby, N. and Zaslavsky, N. (2015). 'Deep learning and the information bottleneck principle', in *2015 IEEE Information Theory Workshop (ITW)* 1–5, doi: 10.1109/ITW.2015.7133169.

Tishby, N., Pereira, F. C., and Bialek, W. (1999). 'The Information Bottleneck method', in *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*.

Tucker, M., Levy, R. P., Shah, J., and Zaslavsky, N. (2022a). 'Generalization and translatability in emergent communication via informational constraints', in *NeurIPS 2022 Workshop on Information-Theoretic Principles in Cognitive Systems*, https://openreview.net/forum?id=yf8suFtNZ5v.

Tucker, M., Levy, R. P., Shah, J., and Zaslavsky, N. (2022b). 'Trading off utility, informativeness, and complexity in emergent communication', in A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho (eds.), *Advances in Neural Information Processing Systems*, https://openreview.net/forum?id=O5arhQvBdH.

Twomey, C. R., Roberts, G., Brainard, D. H., and Plotkin, J. B. (2021). 'What we talk about when we talk about colors', *Proceedings of the National Academy of Sciences* 1180(39): e2109237118.

Uegaki, W. (2022). 'The informativeness/complexity trade-off in the domain of Boolean connectives', *Linguistic Inquiry*, 1–39.

Uegaki, W., Mucha, A., Imel, N., and Steinert-Threlkeld, S. (2023). 'Deontic priority in the lexicalization of impossibility modals'.

von der Gabelentz, G. (1901). *Die Sprachwissenschaft. Ihre Aufgaben, Methoden und bisherigen Ergebnisse*. Leipzig: Chr. Herm. Tauchnitz.

9780192886491_Book.indb 439 27-Mar-25 23:04:45

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Xu, Y., Regier, T., and Malt, B. C. (2016). 'Historical semantic chaining and efficient communication: The case of container names', *Cognitive Science* 400(8): 2081–2094.

Xu, Y., Liu, E., and Regier, T. (2020). 'Numeral systems across languages support efficient communication: From approximate numerosity to recursion', *Open Mind* 4: 57–70.

Zaslavsky, N. (2020). *Information-Theoretic Principles in the Evolution of Semantic Systems*. PhD Thesis, The Hebrew University of Jerusalem.

Zaslavsky, N., Kemp, C., Regier, T., and Tishby, N. (2018). 'Efficient compression in color naming and its evolution', *Proceedings of the National Academy of Sciences* 1150(31): 7937–7942, doi: 10.1073/pnas.1800521115.

Zaslavsky, N., Regier, T., Tishby, N., and Kemp, C. (2019). 'Semantic categories of artifacts and animals reflect efficient coding', in *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*.

Zaslavsky, N., Kemp, C., Tishby, N., and Regier, T. (2020). 'Communicative need in colour naming', *Cognitive Neuropsychology* 370(5-6): 312–324.

Zaslavsky, N., Hu, J., and Levy, R. (2021a). 'A rate–distortion view of human pragmatic reasoning', *Proceedings of the Society for Computation in Linguistics* 4, doi: 10.7275/gc1z-ck09.

Zaslavsky, N., Maldonado, M., and Culbertson, J. (2021b). 'Let's talk (efficiently) about us: Person systems achieve near-optimal compression', in *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*.

Zaslavsky, N., Garvin, K., Kemp, C., et al. (2022). 'The evolution of color naming reflects pressure for efficiency: Evidence from the recent past', *Journal of Language Evolution*: 04. ISSN 2058-458X. doi: 10.1093/jole/lzac001. lzac001.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley Press.

9780192886491_Book.indb 440 27-Mar-25 23:04:45

CHAPTER 23

CAUSAL GRAPHS AS A TOOL FOR EXPLORING LANGUAGE EVOLUTION

KATERYNA KRYKONIUK AND SEÁN G. ROBERTS

23.1 Introduction: What is the method of causal inference?

Causal inference is a relatively recent method for identifying causal relationships from observational data (see Pearl 1995; Glymour et al. 2016). While the maxim of 'correlation does not equal causation' is important when looking at the relationships of two variables in isolation, it turns out, rather surprisingly, that some conclusions about causality can be made by combining standard probability theory and hypotheses about how a network of multiple variables affect each other. Causal inference has its roots in path analysis, which was itself invented to study genetic evolution (see Wright 1983), and has recently been expanded in the field of artificial intelligence, mathematics, and epidemiology, notably by Judea Pearl and colleagues (Pearl 1995, 2010; Pearl and Mackenzie 2018). Causal inference is associated with causal graphs: a tool for precisely expressing causal hypotheses in a graphical format. Not only does it allow researchers to communicate their ideas clearly and in a structured way, but also provides access to powerful mathematical insights for synthesizing theories and designing empirical work. For example, it can be used during review to summarize and synthesize theories in the literature, during hypothesis formulation, during experiment design to identify variables that need to be measured or manipulated, during analysis to help define statistical models, and during the communication of research findings via graphical representations.

Since the field of language evolution involves synthesizing complex hypotheses from many different disciplines and lines of evidence, we argue that causal graphs can be a critical tool for planning empirical studies and developing theory. Hence,

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

this chapter considers the method of causal graphs in greater detail. In section 23.2, we first look at its fundamental theoretical concepts, assumptions, and limitations, illustrating them with real-world examples. Then, section 23.3 identifies areas of research that may benefit from the application of this method and presents the Causal Hypotheses in Evolutionary Linguistics Database (CHIELD). We illustrate how it is possible to get valuable insights from applying causal graph methodology to the studies on the emergence of symbolic signalling and from synthesizing different archaeological and anthropological findings. Finally, section 23.4 outlines what we have learnt from the method of causal inference thus far, and section 23.5 concludes this chapter.

23.2 The fundamentals of causal graphs

Causal graphs are visual, mathematical expressions of causal relations. They represent hypotheses about causal relations, not proof of hypotheses, and allow us to create a holistic picture of the studied phenomenon and trace the relationship between variables. Causal graphs are backed up by Pearl's (1995) *do*-calculus: they hinge on a set of key concepts and assumptions which we consider closely in this section.

The basic elements of causal graphs are nodes and edges. Nodes (or 'vertices') represent measurable variables which can be discrete (Bellot et al. 2022) or continuous (Zhang and Bareinboim 2021). Edges (or arrows) in causal graphs signify a causal relationship: the counterfactual claim that if the source variable were to change sufficiently, then the destination variable would change. This relationship might be probabilistic (statistical) or deterministic (absolute).

Let us illustrate the basics of causal graphs with a simple, intuitive example (we will address language evolution theories later in this chapter). We might observe that when someone goes to the beach, they often come back with stains on their shirt. We imagine a hypothesis that this is because of the ice cream they bought. Figure 23.1 shows a hypothesis about the relationship between three variables: whether someone is at the beach, whether they have bought an ice cream, and whether they have a stain on their shirt. The arrows between the variables hypothesize two causal effects: first, if you were to change your location from somewhere else to the beach, your probability of buying an ice cream would change (it would increase, though that is not represented in the graph). Second, if you were to buy an ice cream, the probability of having a stain would increase. In this graph, there is no arrow in the opposite direction, so if the stain variable were to change, this would not affect the probability of buying an ice cream. Similarly, there is no direct arrow from the beach variable to the stain variable, which represents the hypothesis that there is no causal effect of being at the beach on shirt stains, except for the effect that it might have on you buying an ice cream. The one-directionality of the causal flow means that the graph in Figure 23.1 is 'acyclic'.

9780192886491_Book.indb 442 27-Mar-25 23:04:45

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN



FIGURE 23.1 The basic elements of causal graphs: nodes and edges.

Causal connections can be represented with the formal notation of conditional probability. Assuming that, for a given person at a particular time of observation, each variable is a binary truth state (either *true* or *false*), we can write:

$$\mathrm{P}(stain \mid icecream) > \mathrm{P}(stain)$$

That is, suppose that we surveyed a hundred people at the beach. The probability of observing someone with a stain, given that they have bought an ice cream, is higher than the overall probability of having a stain. However, this model of the world is incomplete, since stains may be caused by some other factors—for example, children are messy. To distinguish an observational probability from a causal relationship, causal graph theory introduces the *do*-operator which implies intervention:

$$\mathrm{P}(stain \mid \mathrm{do}(icecream)) > \mathrm{P}(stain)$$

This means that, if we were to intervene and give someone an ice cream, the probability of them getting a stain would increase. Working out this probability might require a formal experiment, but, importantly, we can now formally distinguish observation and intervention. This is the foundation of *do*-calculus (Pearl 1995) that eventually lets us make statements about intervention just from observations.

However, to fully engage with the *do*-calculus, we still need to consider a few more concepts: 'confounding', 'confounder', 'back-door criterion', and 'collider'. They are important for understanding the structure of causal graphs, checking the assumptions of causality, and identifying the design of an experiment or hypothesis.

The first key concept is the 'back-door criterion'. As we have already established, causality flows in one direction. However, correlations can appear between variables connected in any direction. From observational data alone (e.g. *ice cream* and *stain*), it is not evident whether the nature of this relation is causal or correlational until we evaluate these events in a more general context of other variables. To illustrate, let us explore a more sophisticated causal model (Figure 23.2, left). Here, we hypothesize that younger people are more likely to have stains in general, and they are also more likely to go to the beach. That is, there is a causal link between *age* and *stain*, and between *age* and *beach*. If this model of the world is correct, it represents a confound (*age* is a 'confounder'): we might expect that some of the correlation between *beach* and *stain* is due to the common cause of *age*, rather than the causal path through *ice cream*. Put another way, we could expect to observe some correlation between *beach* and *stain* even if there was no causal effect *beach* → *ice cream* → *stain*.

9780192886491_Book.indb 443 27-Mar-25 23:04:46

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

This is formally represented in the structure of the graph: there are two paths from the variable *go to the beach* to the variable *shirt stain*: one following the causal arrows (solid arrows, *Go to the beach* → *Buy Ice cream* → *Shirt stain*), and one following a 'non-causal' path (in Figure 23.2, highlighted with a dashed line). The latter is the 'back-door criterion' or 'back-door' path.

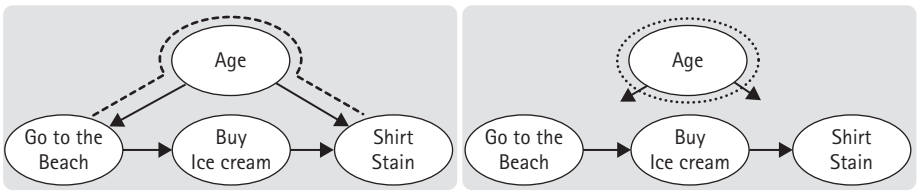


FIGURE 23.2 An example of a causal model: (left) with the 'back-door' correlation path (represented with a dashed line); (right) breaking this path by controlling for variable age (represented with a dotted line).

If we want to estimate the causal link from *beach* to *stain*, we need to 'intervene' such that the alternative 'back-door' path is blocked. In our example model (Figure 23.2, right), this is possible by controlling for the variable *age*: we can run a randomized controlled experiment by randomly assigning people to go to the beach (removing the causal effect *age* → *beach*); or we can only look at people of a specific age (removing *age* → *beach* and *age* → *stain*); or we can control for the effect of *age* in a statistical model.

Most researchers can intuitively identify confounds and variables that need to be controlled for. However, this task becomes more difficult in more complex models. Declarative and procedural approaches to identifying key variables may be incomplete or inaccurate (Pearl and Mackenzie 2018). The causal perspective, on the other hand, provides simpler and more precise solutions to this problem by giving clear instructions on how to detect a confounder. As we have already learnt, confounders are any variables that cause a difference between the probability from an observation and the probability from an intervention, and they can be inferred from the structure of the causal graph (Shrier and Platt 2008). On the causal path of X → Y, a confounder (i) has a causal arrow into X; (ii) is located on a non-causal, 'back-door' path from X to Y; and (iii) is not a descendent of X (i.e. is not causally affected by X, neither directly nor indirectly). In Figure 23.2 (left), the variable *age* is a confounder: it points to *stain*; it belongs to the non-causal, 'back-door' path from *beach* to *stain*; and it is not causally affected by *beach*.

Confounders create spurious correlations if they are not controlled for. However, some variables cause spurious correlations if they *are* controlled for. In Figure 23.2, there is a 'collider' structure: *ice cream* → *stain* ← *age*. It has the unintuitive properties of blocking a non-causal path normally, but opening a non-causal path when the central variable is controlled for. To get a sense of this, imagine that we only surveyed people with a shirt stain (effectively controlling for *stain*). If a person with a shirt stain is an adult, they are likely to have grown out of general messiness, which suggests that their stain is more likely to have been caused by an ice cream. Therefore, controlling for the

9780192886491_Book.indb 444 27-Mar-25 23:04:47

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

central variable *stain* in the collider *ice cream* → *stain* ← *age* creates a negative correlation[1] between the two remaining variables.

Given this rather disturbing knowledge, great care is needed when designing experiments and statistical tests. Causal graphs provide a systematic way of making these design decisions. They can also help decide whether an experiment is necessary at all. To demonstrate this, Figure 23.3 shows a more complex causal model with some additional notation. The line with a bar indicates that getting a shirt stain does not causally affect one's age. This is already implied by the absence of a causal arrow in that direction, but it is sometimes useful to explicitly represent the hypothesis. The line ending in a dot represents an indicator relationship: *ethicalness* is a latent variable that can be potentially measured by veganism. The effect of *ice cream* on *energy* has also been added.

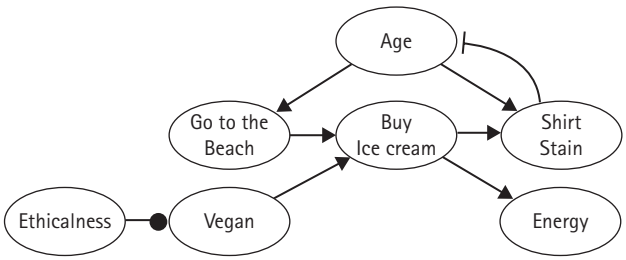


FIGURE 23.3 Other conventions of causal graphs: the line with a bar and round-ended line.

To estimate causal effects from observational probabilities, we can attempt to apply the *do*-calculus (Pearl 1995). This is a set of rules that allow exchanging elements in a probability formula, similar to how variables can cancel each other out in algebra. They let us manipulate causal variables, make models simpler and computationally more efficient, and avoid unnecessary complications and errors. The *do*-calculus is complete and can find an estimate of causal effects based on observations if such a solution exists (Shpitser 2008). If no solution is feasible with the *do*-calculus, an experiment must be conducted instead.

The *do*-calculus operates on three rules that we present below (for more detail, see Pearl 1995; Pearl and Mackenzie 2018). First, if we want to estimate the causal effect of X on Y, the first rule lets us remove a variable Z from an equation on the condition that Y is independent of Z. (i.e. $[Y \perp Z \mid X]$ in a graph $G_{\bar{X}}$ that is a subgraph of the full graph where all arrows into X have been deleted, Pearl 1995: 676):

$$P(Y \mid do(X), Z) = P(Y \mid do(X)) \quad \text{if } (Y \perp Z \mid X)_{G_{\bar{X}}} \tag{1}$$

[1] If we plot *age* on the x-axis and *buy ice cream* on the y-axis, we will observe a negative correlation (going from the left upper corner of this scatterplot to its right bottom corner), since our sample is not random but selected on the basis of people having a shirt stain. That is, we will see a larger number of younger people and a smaller number of elderly people buying ice cream, whereas, in truly random sampling, the scatterplot will approximate a horizontal line.

9780192886491_Book.indb 445 27-Mar-25 23:04:48

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Imagine we wanted to calculate the probability of having a stain if (a) you were on the beach and (b) we intervened to give you an ice cream. According to the rule above, the probability of being on the beach is irrelevant and therefore can be estimated from the intervention alone.

$$\mathrm{P}(\mathit{stain} \mid \mathrm{do}(\mathit{icecream}), \mathit{beach}) = \mathrm{P}(\mathit{stain} \mid \mathrm{do}(\mathit{icecream})$$

That is, this rule allows us to reduce the complexity of a causal model.

The second rule allows us to treat interventions as observations and delete nodes that lie on the main intervention–outcome path. This is possible if Y is independent of Z, given W and X, and if all the incoming arrows to X and all the outgoing arrows from Z are deleted (Pearl 1995: 677):

$$(2) \qquad \mathrm{P}(\mathrm{Y} \mid do(\mathrm{Z}), do(\mathrm{X}), \mathrm{W}) = \mathrm{P}(\mathrm{Y} \mid \mathrm{Z}, do(\mathrm{X}), \mathrm{W}) \qquad \text{if} (\mathrm{Y} \perp \mathrm{Z} \mid \mathrm{W}, \mathrm{X})_{G_{\bar{X}, \underline{Z}}}$$

In Figure 23.3, we are interested in the probability of *stain* given *ice cream*, and suppose we intervened to make someone vegan. Since veganism does not affect any other part of the model, we can estimate the intervention using the observational probability:

$$\mathrm{P}(\mathit{stain} \mid \mathrm{do}(\mathit{vegan}), \mathit{icecream}) = \mathrm{P}(\mathit{stain} \mid \mathit{vegan}, \mathit{icecream}).$$

Because intervention requires time and effort on the part of the experimenter, this rule enables us to exchange intervention with observation and, thus, save our resources.

Finally, according to the third rule, we can delete interventions if Y is independent of Z, if W and X are given, and that a Z-node is not an ancestor of a W-node in $G_{\bar{X}}$ (i.e. in a modified graph, Z does not precede W):

$$(3) \qquad \mathrm{P}(\mathrm{Y} \mid do(\mathrm{Z}), do(\mathrm{X}), \mathrm{W}) = \mathrm{P}(\mathrm{Y} \mid do(\mathrm{X}), \mathrm{W}) \qquad \text{if} (\mathrm{Y} \perp \mathrm{Z} \mid \mathrm{W}, \mathrm{X})_{G_{\bar{X}, \overline{Z(W)}}}$$

To illustrate the third rule, let us again consider the causal path *go to the beach* → *buy ice cream* → *shirt stain* and the causal arrow from *y ice cream* to *energy* in Figure 23.3. According to the third rule, we can delete an intervention on the *energy* variable (e.g. an experiment where half the participants eat a sugar cube), since this will not have an impact on the model's outcome:

$$\mathrm{P}(\mathit{stain} \mid \mathit{icecream}, \mathrm{do}\ (\mathit{energy})) = \mathrm{P}(\mathit{stain} \mid \mathit{icecream})$$

The *do*-calculus rules can be applied in sequence alongside colliders and other properties of causal graphs to allow us to estimate probabilities from an experimental intervention using probabilities based on observations. For example, though it seems like black magic, it is possible to estimate the causal effect of going

9780192886491_Book.indb 446 27-Mar-25 23:04:50

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

to the beach on shirt stains while controlling for the confound of *age* even if we had no data on age.[2]

These advantages come with some limitations: first, the conclusions reached using the causal axioms depends on the accuracy of the assumptions about the causal relationships. However, while it is very difficult to know whether one's causal model is 'correct', the main use of causal graphs is to develop and express logically consistent hypothetical models from a set of assumptions, not necessarily to represent proof of how the world works. Second, there are challenges when variables influence each other bidirectionally, as is often the case with co-evolving variables. For example, imagine that we have two variables that influence each other at different points in time (e.g. a word that is used frequently might become shorter, and shorter words might be more likely to be selected for use). A causal model $A \rightleftarrows B$ violates the 'acyclic' assumption that underlies many of the causal graph calculus. To overcome this limitation, it is sometimes possible to expand these nodes to represent these variables at different points in time ($A_{t=0} \rightarrow B_{t=1} \rightarrow A_{t=2} \rightarrow B_{t=3} \ldots$). Using this formulation, researchers can be explicit about the order of cause and effect over time and, as a result, the graph remains acyclic.

Further, identifying confounds and independence sets is easy in simple models, but this task becomes more difficult with complex models. For this reason, various algorithms have been developed that discover solutions to causal problems efficiently (e.g. the online tool DAGitty; http://www.dagitty.net/).

23.3 Using causal graphs in evolutionary linguistics

Causal graphs provide four advantages for researchers in the field of language evolution. First, they provide a way of explicitly representing hypotheses. Second, they allow researchers to compare evidence from multiple studies. Third, they can act as a locus for identifying terminological and theoretical differences. And finally, they help researchers synthesize and extend theories. In this section, we illustrate these advantages with various examples taken from the Causal Hypotheses in Evolutionary Linguistics Database (CHIELD; Roberts et al. 2020). CHIELD is a multidisciplinary, open-source database of hypotheses related to language evolution which have been coded manually as causal graphs. Users can discover links between hypotheses, identify conflicting evidence, and spot weak links that currently have little support.

[2] Estimating *beach* → *ice cream* requires blocking the back-door path via *age*. We do not have data on age, but there is already a collider that blocks this path in *ice cream* → *stain* ← *age*. Next, estimating *ice cream* → *stain* also has a back-door path via *age*. But this can be blocked by statistically controlling for *beach*. The two estimates can be combined to calculate the total causal effect.

9780192886491_Book.indb 447 27-Mar-25 23:04:50

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

A causal graph can be considered in its entirety and also be broken down into individual causal connections that might be analysed and tested individually. To illustrate this, Figure 23.4 shows a simplified version of the work of Limor Raviv and colleagues on the connections between population structure and language learning using several different methods. For example, individuals are less likely to share common ground with interlocutors in a sparser network, so may select systematic language (Raviv et al. 2020). However, Raviv et al. (2019) show that population size itself is a direct contributor to systematicity independently of connectivity via the rate of innovation and change. Various studies also show effect of population size and structure on the variability in signals (e.g. Raviv et al. 2022). Finally, systematicity and variability contribute to various measures of language learnability (Raviv et al. 2021). These studies use a number of different methods and sources of evidence but can be usefully brought together to represent the larger theory. For example, the hypotheses suggest that both size and connectivity should be considered when predicting systematicity. Furthermore, there is a potential back-door path from population size, via population structure and connectivity, to systematicity. This might imply that the correlation between population size and systematicity found in Lupyan and Dale (2010) may be a product of both size and connectivity. Establishing the exact dependencies between size, connectivity, age structure, and gender ratio will be key to solving this confound.

The second advantage is that strands of evidence can be compared. For example, Figure 23.4 shows four other studies that investigate the same causal links. There is

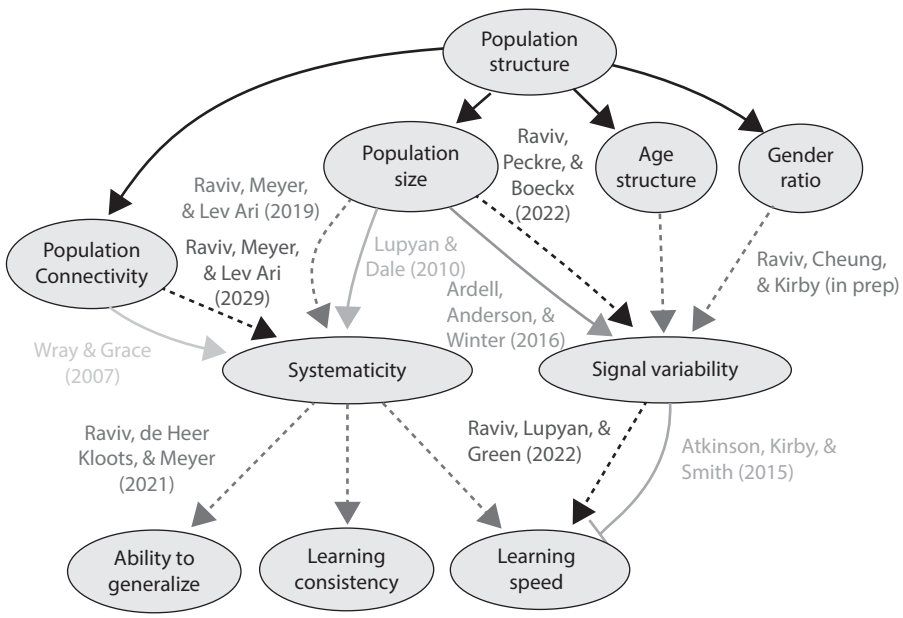


FIGURE 23.4 A causal graph of work by Raviv and colleagues linking population structure to aspects of language learning. Dashed-line arrows indicate studies by Raviv and colleagues, while solid arrows indicate parallel studies.

9780192886491_Book.indb 448 27-Mar-25 23:04:51

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

converging evidence from Wray and Grace (2007), Lupyan and Dale (2010), and Ardell et al. (2016). However, Atkinson et al. (2015) do not find an effect of signal variability on learnability, conflicting with Raviv et al. (2022). The causal graph does not provide an evaluation of these studies, but it highlights differences between results and identifies specific causal links for possible future studies. CHIELD is equipped with a tool for combining causal graphs, making it possible to establish overlapping or conflicting areas between theories, or simply to explore how theories are related to each other.

The value of CHIELD becomes particularly evident when designing empirical studies in evolutionary linguistics that deals with cross-cultural variation of language change. The networks of causal effects in this area are vastly complex and, thus, require a systematic approach to designing statistical models. One of these approaches is path analysis, which combines causal inference with various statistical techniques (e.g. Wright 1984). This method, for example, has been successfully pursed by Josserand et al. (2021) in the study that explains the variation of the colour lexicon of languages with their cultural, biological, and physical environments. CHIELD helps design these kinds of highly interdisciplinary studies: it provides an option for converting causal graphs into R code which can then be used for running path analyses using the *brms* library (Bürkner 2017).

Linking studies in this way depends on each hypothesis using the same vocabulary for the nodes in the graph. When theories come from different methodological grounds, differing terminology is inevitable. However, we argue that causal graphs present an opportunity to identify and engage with terminological differences. For example, in Figure 23.4, it is not entirely clear to us whether all studies have the same conception of systematicity. Indeed, Raviv et al. (2022) review this challenging definition. On CHIELD, anyone can contribute towards creating and editing graphs, and version control tools allow open review and discussion of the data.

Finally, CHIELD can help with theoretical synthesis for studies of earlier stages of language evolution such as pre-adaptation and co-evolution. Since direct observation of the biological evolution of human language is not possible, causal inference provides a systematic way of comparing and evaluating theories. We illustrate this using a synthesis of various theories that aim to explain why humans evolved a symbolic, referential system of communication. Few attempts have been made to consistently synthesize these theories and map them against each other, partly because they are very complex and come from very different areas of study. With the help of CHIELD, Roberts et al. (2022) formalized eleven theories of the evolution of symbolic communication as causal graphs. Figure 23.5 shows a part of the final graph, focusing on Dunbar's (2004) theory of social cognition: the risk of predation creates a need for individuals to build alliances, which increases the group size. Living in larger groups requires individuals to spend more time doing social bonding, like grooming. This factor creates pressure for more efficient bonding strategies, and symbolic communication (gossip) facilitates this process.

9780192886491_Book.indb 449 27-Mar-25 23:04:51

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

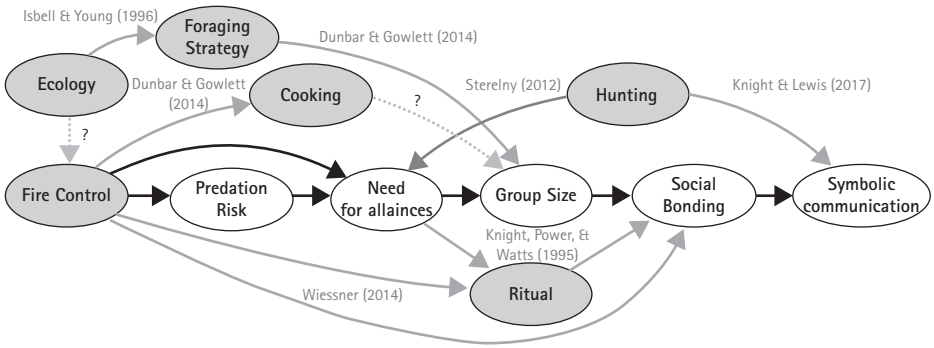


FIGURE 23.5 Dunbar's theory of language evolution (white nodes) in the context of other theories.

Alongside this hypothesis, we map Knight et al.'s (1995) hypothesis which argues that ritual was needed to allow individuals to trust each other enough to believe symbolic signals. While this is often framed as an opposing theory, the two are, from a formal perspective, mutually compatible. Other studies explain the preconditions for Dunbar's hypothesis. For example, the shift of ecology from dense jungle to open savannah and the change in accessibility of resources had an impact on foraging strategies and group size (Isbell and Young 1996). Hunting had a causal effect on the need for alliances (Sterelny 2012) and directly provided a motivation for symbolic signals via the mimicking of animal noises (Knight and Lewis 2017). Similarly, the maintenance and control of fire created logistical challenges that also created a need for alliances (Twomey 2013). In addition, fire was used in ritual, and it extended light into the night to facilitate social bonding (Wiessner 2014). While drawing this graph, we also noted that there are further potential links: use of fire may reduce predation risk, but it also allows for cooking which might support larger group sizes. There are also potential links between fire control, hunting, and foraging, or between the ecology and fire control itself (Wrangham and Conklin-Brittain 2003; Hawkes et al. 2018), which, as far as we know, are not yet incorporated into theories of language evolution. We suggest that understanding this type of interaction will help develop a synthesis between theories and explain how the context of our ancestors provided unique pressures for the emergence of complex communication systems.

23.4 What have we learnt from this method?

The examples discussed above have demonstrated the advantages of causal inference for the study of language evolution. There are four major insights that we can draw: first, theories are more interconnected that we previously thought. Establishing links between theories allows us to create a more complete picture of the problem, and, as a result, we find an opportunity to address the challenge of language development in early

9780192886491_Book.indb 450 27-Mar-25 23:04:51

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

humans on a meta-level. Thus, the realization of theories' interconnectedness is the first step towards the unification of terminology apparatus. Second, many theories in evolutionary linguistics are not mutually exclusive. This suggests there is more work to be done in identifying differences in predictions.

Third, many links in causal graphs have no empirical support. For theories to be valid and convincing, more work needs to be done for recognizing empirically weak areas and for collecting evidence to support them. On the one hand, addressing links with no or little evidence will secure the development of comprehensive theories where *a priori* reasoning and empirical facts converge; on the other, this endeavour has a potential to shape the field of language evolution by directing scientific interests towards important problems that are contextualized within a rich network of arguments. In this view, we are focusing on immediate and pivotal questions.

Finally, the emergence and evolution of language is complex. One way to make progress is to split hypotheses into links that can be tested individually. The visual and formal tools provided by causal graphs make this effort much easier.

For these reasons, we suggest that causal graphs can become a central tool for supporting different stages of research in evolutionary linguistics. For example, we have integrated them into our typical workflow: we use CHIELD to formalize prior research hypotheses and combine them into a synthesized graph. Then we use the causal axioms to eliminate causally irrelevant variables to focus on the effects we are interested in. We have found that drawing causal graphs during collaborative research design has helped us convert abstract concepts into concrete, measurable variables. This is particularly invaluable for experimental semiotics where abstract concepts from the theory must be grounded in specific analogues in the experiment. Furthermore, representations of causal graphs have been very helpful when designing statistical models and communicating our designs in papers.

23.5 Concluding remarks

This chapter introduced causal graphs as applied to the field of evolutionary linguistics, discussing their advantages and shortcomings. Causal graphs can be useful at every stage of research in evolutionary linguistics, including literature review, hypothesis formulation, experiment design, and the communication of findings. CHIELD was discussed as a tool for facilitating these steps. We predict that this methodology will become increasingly important in evolutionary linguistics.

References

Ardell, D., Anderson, N., and Winter, B. (2016). 'Noise in phonology affects encoding strategies in morphology', in S. G. Roberts, C. Cuskley, L. McCrohon, et al. (eds.), *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANGX11)*.

9780192886491_Book.indb 451 27-Mar-25 23:04:51

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Atkinson, M., Kirby, S., and Smith, K. (2015). 'Speaker input variability does not explain why larger populations have simpler languages', *PloS One* 10(6): e0129463.

Bellot, A., Zhang, J., and Bareinboim, E. (2022). *Scores for Learning Discrete Causal Graphs with Unobserved Confounders*. Columbia Causal AI Laboratory, Technical Report (R-83).

Bürkner, P. (2017). 'brms: An R package for Bayesian multilevel models using Stan', *Journal of Statistical Software* 80(1): 1–28.

Dunbar, R. I. (2004). 'Gossip in evolutionary perspective', *Review of General Psychology* 8(2): 100–110.

Glymour, M., Pearl, J., and Jewell, N. P. (2016). *Causal Inference in Statistics: A Primer*. New Jersey: John Wiley & Sons.

Hawkes, K., O'Connell, J., and Blurton Jones, N. (2018). 'Hunter-gatherer studies and human evolution: A very selective review', *American Journal of Physical Anthropology* 165(4): 777–800.

Isbell, L. A. and Young, T. (1996). 'The evolution of bipedalism in hominids and reduced group size in chimpanzees', *Journal of Human Evolution* 30(5): 389–397.

Josserand, M., Meeussen, E., Majid, A., and Dediu, D. (2021). 'Environment and culture shape both the colour lexicon and the genetics of colour perception', *Scientific Reports* 11(1): 1–11.

Knight, C. and Lewis, J. (2017). 'Wild voices: Mimicry, reversal, metaphor, and the emergence of language', *Current Anthropology* 58(4): 435–453.

Knight, C., Power, C., and Watts, I. (1995). 'The human symbolic revolution: A Darwinian account', *Cambridge Archaeological Journal* 5(1): 75–114.

Lupyan, G. and Dale, R. (2010). 'Language structure is partly determined by social structure', *PLOS ONE* 5(1): e8559, https://doi.org/10.1371/journal.pone.0008559.

Pearl, J. (1995). 'Causal diagrams for empirical research', *Biometrika 82*(4): 669–688.

Pearl, J. (2010). 'The foundations of causal inference', *Sociological Methodology* 40(1): 75–149.

Pearl, J. and Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books.

Raviv, L., Meyer, A., and Lev-Ari, S. (2019). 'Larger communities create more systematic languages', *Proceedings of the Royal Society B* 286(1907): 20191262.

Raviv, L., Meyer, A., and Lev-Ari, S. (2020). 'The role of social network structure in the emergence of linguistic structure', *Cognitive Science* 44(8): e12876.

Raviv, L., de Heer Kloots, M., and Meyer, A. (2021). 'What makes a language easy to learn? A preregistered study on how systematic structure and community size affect language learnability', *Cognition* 210: 104620.

Raviv, L., Lupyan, G., and Green, S. C. (2022). 'How variability shapes learning and generalization', *Trends in Cognitive Sciences* 26(6): 462–483.

Raviv, L., Peckre, L. R., and Boeckx, C. (2022). 'What is simple is actually quite complex: A critical note on terminology in the domain of language and communication', *Journal of Comparative Psychology* 136(4): 215–220.

Roberts, S. G., Killin, A., Deb, A., et al. (2020). 'CHIELD: The causal hypotheses in Evolutionary Linguistics Database', *Journal of Language Evolution* 5: 2, https://doi.org/10.1093/jole/lzaa001.

Roberts, S. G., Irvine, E., and Jordan, F. M. (2022). 'The common task framework: Using causal theories from linguistic anthropology to explore the emergence of symbolic communication', in A. Ravignani, R. Asano, D. Valente, et al. (eds.), *Proceedings of the Joint Conference on Language Evolution (JCoLE)*, https://doi.org/10.17617/2.3398549.

9780192886491_Book.indb 452 27-Mar-25 23:04:51

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Shpitser, I. (2008). *Complete Identification Methods for Causal Inference* (Doctoral dissertation, UCLA).

Shrier, I. and Platt, R. W. (2008). 'Reducing bias through directed acyclic graphs', *BMC Medical Research Methodology* 8(1): 1–15.

Sterelny, K. (2012). 'Language, gesture, skill: The co-evolutionary foundations of language', *Philosophical Transactions of the Royal Society B* 367(1599): 2141–2151.

Textor, J., Hardt, J., and Knüppel, S. (2011). 'DAGitty: A graphical tool for analyzing causal diagrams', *Epidemiology* 22(5): 745, http://www.dagitty.net/

Twomey, T. (2013). 'The cognitive implications of controlled fire use by early humans', *Cambridge Archaeological Journal* 23(1): 113–128.

Wiessner, P. W. (2014). 'Embers of society: Firelight talk among the Ju/'hoansi Bushmen', *Proceedings of the National Academy of Sciences* 111(39): 14027–14035.

Wrangham, R. W. and Conklin-Brittain, N. (2003). 'Cooking as a biological trait', *Comparative Biochemistry and Physiology* 136: 35–46.

Wray, A. and Grace, G. W. (2007). 'The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form', *Lingua* (117): 543–578.

Wright, S. (1983). 'On "Path analysis in genetic epidemiology: A critique" ', *American Journal of Human Genetics* 35: 757–768.

Zhang, J. and Bareinboim, E. (2021). 'Bounding causal effects on continuous outcomes', in *Proceedings of the 35th AAAI Conference on Artificial Intelligence*. Columbia Causal AI Laboratory, Technical Report (R-61).

9780192886491_Book.indb 453 27-Mar-25 23:04:51









OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

PART III

ANIMAL DATA

9780192886491_Book.indb 455 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:04:51

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

9780192886491_Book.indb 456 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:04:51

CHAPTER 24

OBSERVATIONAL WORK

MARCO GAMBA AND CHIARA DE GREGORIO

Observing animal or human subjects makes it possible to understand how often particular behaviours occur, with what internal or external stimuli they are associated with, and what variation certain displays undergo with the context of emission. Observing behaviour is fundamental for formulating hypotheses on welfare and defining and interpreting manipulation experiments. There has been a growing interest in conducting comparative animal studies. Indeed, we have witnessed a widespread trend to test hypotheses about critical elements of language evolution in different animal species. Nevertheless, unfortunately, this large body of work is based mainly on observational studies, which are sometimes undesirable for researchers not trained in biological, natural, or psychological sciences.

It is impossible to cover exhaustively in one chapter such a broad topic as the issues related to observational studies (for which we suggest other excellent sources such as Bateson and Martin 2021). This chapter aims to briefly summarize some of the 'hot topics' concerning the choices a student or researcher has to make before starting a study of animal behaviour from the perspective of the evolution of language faculties. As a theoretical-only treatment would be possibly boring for many, we also decided to add some subjective practical suggestions, also potentially useful for those approaching these kinds of studies for the first time.

If observing the behaviour of an animal is one of the simplest and more direct methods to gain information about its ethology and ecology, as it could be virtually done with no equipment other than pen and paper, technological development has strongly influenced observational animal studies like many other aspects of biological and naturalistic investigations. For example, advanced recording devices, robust storage systems, and long-lasting batteries have profoundly changed ethological data collection approaches. Similarly, data processing is now linked to an exquisitely computerized use of sampling and often involves analyses based on machine learning or sophisticated statistical approaches.

The new challenges arising from these studies have prompted new software development, often based on open-source technologies, which sometimes show interoperability between different platforms and systems.

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

24.1 'A GOOD START IS HALF THE BATTLE'

Starting with an appropriate sampling rule is critical for an observational study. There are many factors that can influence the adoption of a particular method, such as the number of individuals that we want to observe at the same time, their visibility, the time window that we are interested in, and the scale of the behaviour that we want to analyse.

During *Ad libitum sampling*, the observer freely records whatever is visible and possibly indulges in the behaviours occurring more frequently. This sampling rule is common during preliminary observations, especially when completing an ethogram (the list of behaviours you will observe in a systematic series of observations).

Focal sampling is possibly the most common and the most helpful sampling rule. During focal observation, the observer collects the behaviours of an individual for a specified amount of time, attributing each action or state to a record of predetermined behaviours. Ideally, observations of the individuals should follow a randomized and systematic schedule that allows sampling a particular subject at different times of day and during different sessions. We must record and quantify if the focal subject becomes out of sight during observation.

For instance, Spezie and colleagues (2023) used focal animal sampling when investigating the response of indris (*Indri indri*; De Gregorio et al. 2022a) to the simulated presence of intruders in their territory. They broadcasted the songs of familiar and non-familiar social groups: the focal indris were followed for at least one hour before the playback and behavioural responses were noted before and after the onset of the playback. Each animal in the group was followed by a different observer to collect the behaviour of all group members. They found that indris responded more rapidly and approached the speaker more often in response to playback stimuli of non-familiar individuals than to stimuli of neighbouring groups. These lemurs were able to discriminate between different classes of intruders based on the acoustic features of their songs. Observations are often combined with other investigation methods, such as recording of vocal behaviour. Focal animal sampling has also been used in the examination of semantic properties of the alarm calls of vervet monkeys (*Chlorocebus pygerythrus*): in addition to recording monkeys' vocalizations, researchers conducted focal animal sampling of the behaviours preceding and following each utterance and found that different alarm calls were emitted in the presence of different types of predators (Seyfarth et al. 1980), but a more recent analysis showed that there is some gradation between calls emitted in aggressive and alarm contexts (Price et al. 2015).

The focal animal sampling technique can also help the investigation of linguistic laws in species other than our own: for example, Valente and colleagues (2021) and Heesen and colleagues (2019) used focal sampling to collect data, respectively, on the vocal communication of our distant relatives (lemurs) and on the gestural communication of our closest ones (chimpanzees). They found that, in both cases, the communicative repertoire conformed to the Menzerath–Altmann law, which predicts that the compensation between the duration of a construct and that of its components, and to Zipf's law

9780192886491_Book.indb 458 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:04:51

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

of abbreviation, which posits an inverse relationship between signal duration and its occurrence.

These works showed how the focal animal sampling technique can be used to highlight the similarities between human and non-human primates in signal production and perception.

Scan sampling is advantageous when the observer has to scan various subjects in a short time (e.g. a social group). In the case of scan sampling, scans can happen continuously or at regular intervals. Therefore, the observer records instantaneous individuals' behaviour at each scan. Although individual recognition is always desirable, focal animal sampling clearly emphasizes that what we can extract from our observations can be clearly labelled according to the investment of a particular individual or the sequence of behaviours that an individual presents. When we approach an observation with scan sampling, the emphasis is placed on the occurrence of behaviour within the group with less accuracy on its timing. Some studies may integrate focal animal sampling and scan sampling, alternating focal observation of a particular group member (e.g. every five minutes) and scan sampling of all group members (e.g. in the next five minutes). We can sample behaviour using a continuous recording, which helps record accurate timing, duration, frequencies, or time sampling, alternating the times we record behaviours and times in which behaviour is not recorded. Scan sampling is a useful approach in investigating group dynamics and coordination in human and non-human animals. For instance, to understand the dynamics of audience applause, Mann and colleagues (2013) observed the time at which people in an audience started and stopped clapping using video footage, and found that the probability of an individual to start clapping increased with the number of other clapping individuals. Similarly, Bousquet and colleagues (2011), investigated how meerkats (*Suricata suricatta*) use vocal signals to coordinate movement decisions. They observed the animals' foraging behaviour, taking note of the occurrence of moving calls from different individuals, and found that the more group members joined the chorus, the higher the probability that the group would move to a different foraging patch.

One–zero sampling is another typical time-sampling approach, which can also be derived from focal animal sampling or scan sampling via software. In one–zero sampling, ones correspond to those analysis windows in which the behaviour is performed, and zeros to those in which it is not displayed. The duration of the analysis window may change from one study to another, depending on the observation's accuracy and duration. Moreover, one–zero sampling has the potential to collect more behaviours during a pre-defined period, as multiple behaviours can be recorded during each interval. Bolt (2020) used both focal and one–zero sampling to investigate how ring-tailed lemurs (*Lemur catta*) regulate social interaction with the use of affiliative contact calls: every 2.5 min interval, she took note on whether or not the focal animal vocalized, on type of emission, on any agonistic interactions involving the focal subject, and on any group-wide travel events. She found that two types of vocalizations (*chirps* and *wails*) promote social cohesion during group travels, pointing out that the complexity of call usage in strepsirrhine primates can shed light on the evolution of social life of early primates.

9780192886491_Book.indb 459 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:04:52

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

All occurrence sampling is an observational method in which researchers select some specific behaviour of interest and record all the occurrences. It is used to get information about the occurrence of a particular behaviour (or set of behaviours) within a group, especially its rate, frequency, or behavioural synchrony. The behaviours must be easily measured to be adequately recorded and not occur too frequently. For example, Valente and colleagues (2023) sampled all the occurrences of yawning behaviour to assess the presence of yawn contagion in the indris (*Indri indri*). By recording the time of day, the identity of the first yawner, the potential responders, and which of these yawned or not within a 3 min time window from the trigger, the authors found that in this species, yawn contagion is present, the first evidence in lemurs.

24.2 Tools of the trade

One of the increasingly frequent features in studying language and language evolution is the quantitative aspect. The inherent variability of behavioural displays may require more individuals to test hypotheses and find empirical support. Consequently, more extended collections will emphasize accumulating more data, storing it, and managing it as much as possible within an agile and productive workflow. Researchers frequently need to code considerable quantities of video and audio recordings. Although many commercial and open-source software are available, we will concentrate on the ones we know best to indicate some practices we found helpful for our research.

Open-source and multiplatform standalone programs allow user-specific coding in a computer-based environment. This framework is valid for both previously recorded videos and live observations. Species-specific approaches are still common, but we mostly find programs that can allow the customization of settings, ethograms, and operations. These user-defined settings can be shared with collaborators or imported from previous projects.

Studies on the evolution of language often include multimodal aspects and, therefore, the processes we can perform on video recordings and the processing of audio signals must be combined. This is particularly efficient when we can view the video footage and the audio recorded simultaneously with that video from the same device or an external device synchronized with the camera. BORIS, free software available thanks to the University of Turin (Friard and Gamba 2016), has been very successful in recent years because it integrates some of the functions mentioned above (Figure 24.1). Furthermore, the variety of export functionalities and the integrated view with other analysis environments have proved particularly popular with behavioural scientists, also, due to the availability of the source code on GitHub, the level of customization is potentially limitless.

BORIS can work via focal animal, all-occurrence, *ad libitum*, and scan sampling (Figure 24.2). Once keys on the keyboard are associated with behaviours, behavioural

9780192886491_Book.indb 460 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:04:52

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

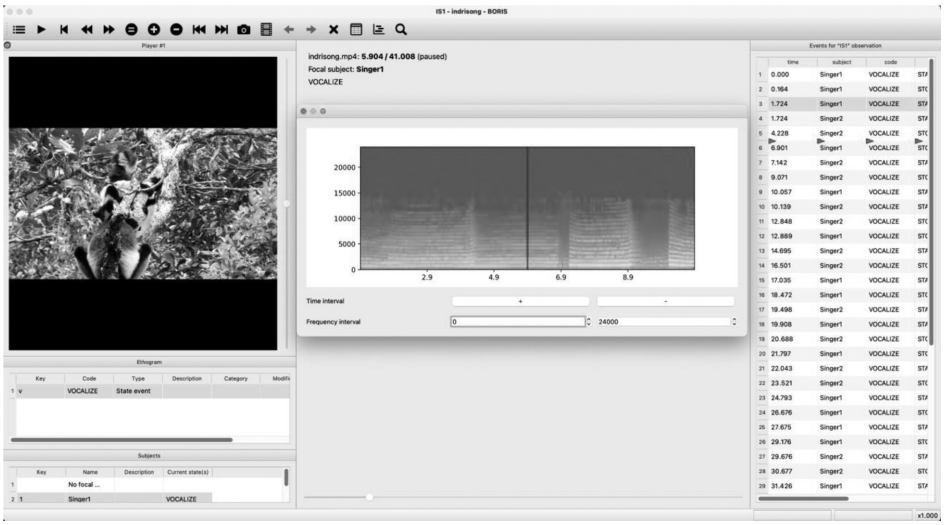


FIGURE 24.1 The BORIS interface. Above is the toolbar with which you can manage the player, slow down or speed up the media playback speed, capture a snapshot, or activate the frame-by-frame function. On the left is the media displayed in Player #1. BORIS can display video, audio, or picture folders; up to eight media can be shown simultaneously. Below the media, you will find the ethogram, including the behaviours to log, and the list of subjects below. The number of behaviours and subjects that can be entered in BORIS is unlimited. In the centre, you can see the information about the media being screened, the point at which you are, the total duration, the focal subject at that particular time, and the current behaviour. Below, optionally, you will find the spectrogram and its display options. Finally, to the right is the coding of the media, with time, subject, and behaviour code.

events can be coded with BORIS by pressing a particular key. Once the key is pressed, BORIS will record the time of the video at which the pressing has happened and will set it as the start of the behaviour (in the case it is a 'state', a behavioural display for which we want to calculate duration) or will indicate the time at which the behaviour has occurred (if it is a 'point event', a behaviour for the duration is not considered). In the case of a 'state', a subsequent pressing of the same key will indicate the stop of that particular behavioural display.

We can test hypotheses on individual or subgroup differences in time allocation by repeating this sequence of operations for many behaviours and across individuals. We can also analyse trends in the progression of behaviours and interactions between subjects using sequential analysis (see Bakeman and Quera 2011). BORIS allows exporting the coding of observations in formats that can be quickly processed for generating time budgets and flow diagrams using open-source programs (e.g. R, Behatrix).

We can use information derived from coding to interoperate with other programs (e.g. Praat: Boersma and Weenink 2020). For instance, the behavioural coding of an observation can be exported in the form of Praat TextGrid and used to segment a long file or extract target labels (Figure 24.3).

9780192886491_Book.indb 461 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:04:52

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

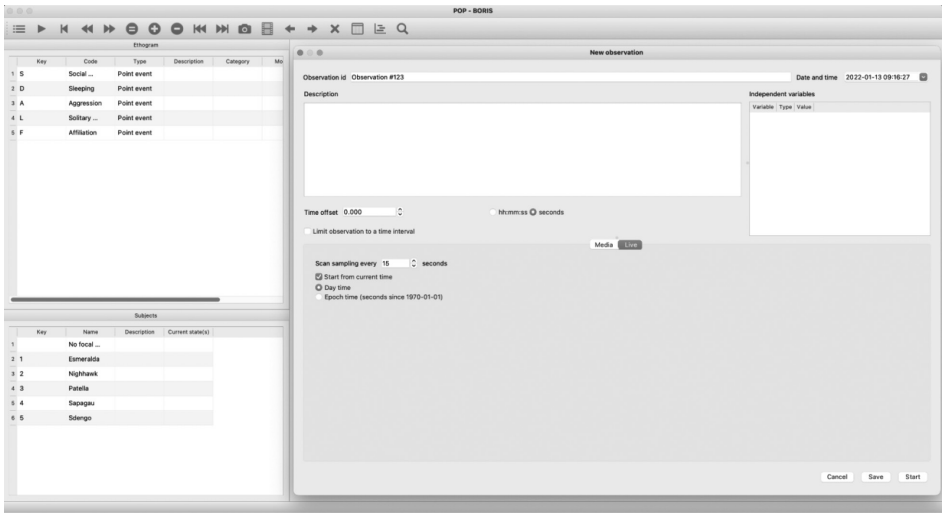


FIGURE 24.2 Scan Sampling options in BORIS. In line with what was once done during manual observations of behaviour, which took place almost exclusively live and with pen and paper, it is possible to organize a scan sampling in which an acoustic signal signals the interval between scans. There is no need for an external beeper. Instead, BORIS will emit an acoustic signal during a live observation when the user-defined interval expires.

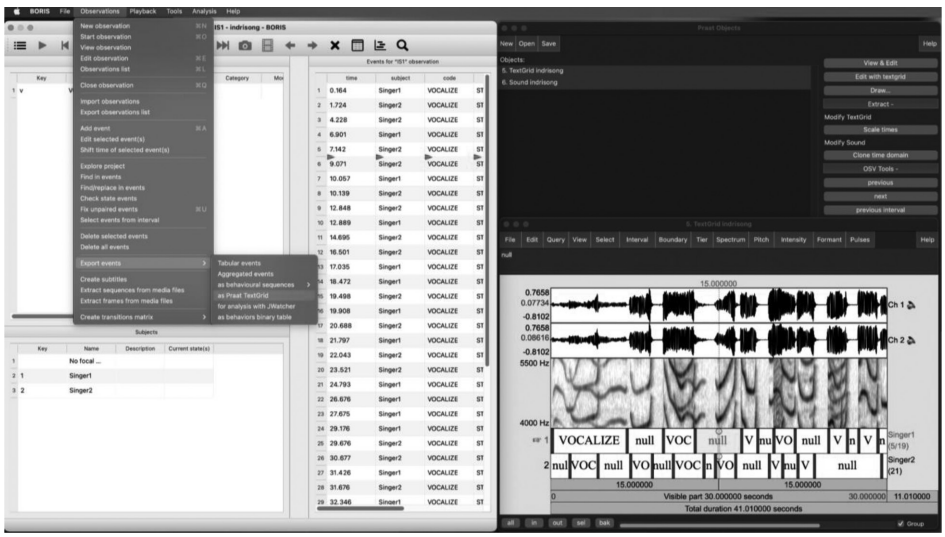


FIGURE 24.3 BORIS can export observations as Praat textgrids. It is easy to export a BORIS observation in Praat textgrid format to extract the parts of interest based on what have become annotations. Extraction can be done manually or by proceeding with a script targeting a particular behaviour or set of behaviours.

9780192886491_Book.indb 462 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:04:52

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

BORIS has been employed in a variety of different studies aimed to describe the communication of human and non-human animals, from dolphins' vocal signals during cooperation tasks (King et al. 2021), to non-verbal communication between pedestrian and drivers at the point of crossing (Rasouli et al. 2017), to investigation of the multimodal nature of red-capped mangabeys' communication (Aychet et al. 2021), and the evaluation of tic suppressibility in patients with Tourette's disorder (McGuire et al. 2022). McGuire and colleagues (2022) used BORIS when investigating tic behaviour in young people with Tourette's disorder, particularly whether the tic suppression capabilities predicted tic severity reductions and treatment response to behaviour therapy. Before administering the behavioural treatment, they coded the video-recorded tic occurrences during two different 5 min tasks, one in which patients could tic freely and the other in which they had to try to suppress the tic while fixating on the dot. The tic frequency in the two tasks was used to calculate each participant's tic suppressibility score. The authors found that baseline tic suppression did not predict treatment outcome, indicating that even young people with difficulty suppressing tics would still benefit from behavioural therapy, exhibiting reductions in tic intensity (McGuire et al. 2022). With the aim to investigate the evolutionary roots of the multimodal nature of human communication, Aychet and colleagues (2021) used BORIS to code all communicative events and the social interactions associated in videos of interacting captive red-capped mangabeys (*Cercocebus torquatus*). They noted down signaller and receiver identities, signal characteristics, and social context. They considered three types of signal categories: body signals, facial signals, and vocal signals, which could be perceived by the receiver as visual, tactile, or audible. Mangabeys frequently produced communication signals in sequences, and multicomponent communication was predominant, especially in contexts of play and aggression. Humans' communication is deeply multimodal (Levinson and Holler 2014) and it frequently combines gestures, facial expressions, and acoustic signals: the work of Aychet and colleagues (2021) highlights the similarities between human and non-human primate communication modalities, and stresses the importance of adopting a multimodal and multicomponent approach to the study of non-human primate communication.

Traditionally, data collected through observation of behaviour point to two types of evaluation. The first concerns the energy and time investment of individuals and is usually visualized with a bar graph in which each bar represents how much a particular behaviour occurred over the total observation time (Figure 24.4). For instance, Scheun and colleagues (2015) used *ad libitum* sampling to compare the behaviour of urban and rural populations of African lesser bushbabies (*Galago moholi*), and found that moving and resting were the activities in which animals spent most time in both conditions. Nevertheless, they also found some differences, in that urban galagos spent considerably more time communicating, feeding, and foraging than the rural population (Scheun et al. 2015). Several adjustments are necessary if many animals are observed at different times from each other. When we derive data from observation, such as using focal animal sampling, we can also construct flow diagrams showing how behaviours follow each other and which transitions occur significantly more often than others. BEHATRIX (Friard et al. 2023) is free and open-source software available for GNU/Linux, Windows, and Mac OS X. By submitting strings of behaviours resulting from a

9780192886491_Book.indb 463 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:04:52

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

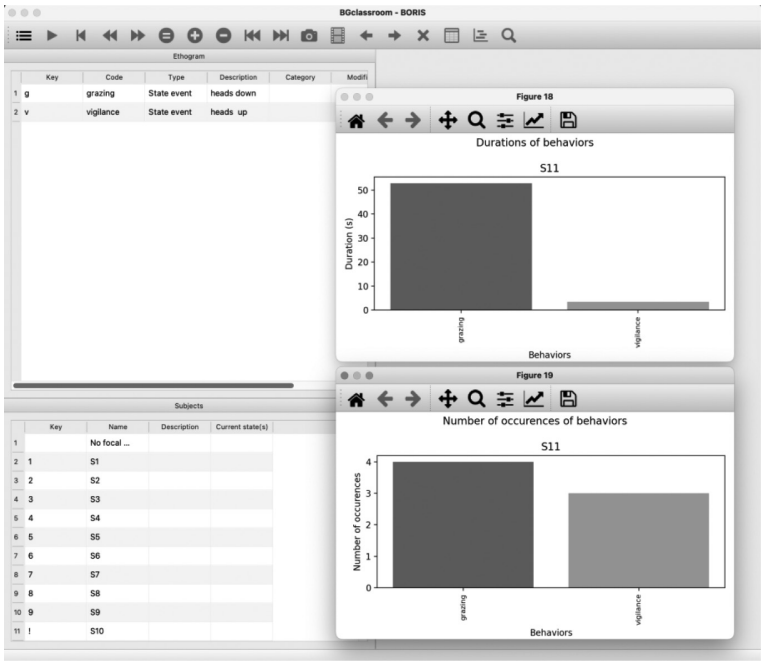


FIGURE 24.4 Time budget generated by BORIS. To visualize your observations at a glance, BORIS can generate bar plots showing the time budget resulting from your analyses. In the figure, from the traditional observation of Grazing and Vigilance in the Barnacle Geese, time budgets of individual S11 are shown for the total duration of Grazing and Vigilance behaviour and the number of occurrences.

previous analysis in BORIS or any other software, BEHATRIX can generate the code for a flowchart representing the transitions between different behaviours and performs a random permutation test to indicate the statistical significance associated with the transitions (Figure 24.5). BEHATRIX was used to investigate turn-taking dynamics in the indris (De Gregorio et al. 2022b), whose powerful songs are used to regulate inter- and intra-group spacing and can be emitted as songs or as choruses. By identifying singers and number of vocalizing animals and annotating when an indri would join the chorus and the exact timing in which it would stop singing, the authors found that the alternation between emitters was not random. In particular, offspring were more likely to drop out from the parents' duet than join in. BEHATRIX can also calculate distances between sequences of behaviours, such as Levenshtein distances (Figure 24.6) and Needleman-Wunsch identities. The Levenshtein distance is a logical distance that measures the distances between two strings of data (e.g. phonetic distance between human dialects; Gooskens and Heeringa 2004), while the Needleman-Wunsch algorithm computes the optimal global alignment between two sequences (e.g. identification of linguistic macrofamilies given language similarities; Jäger 2015). Zanoli and colleagues (2020) quantified variability in phrase concatenation using the Levenshtein distance. The authors analysed the degree of similarity within and between individuals

9780192886491_Book.indb 464 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:04:52

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

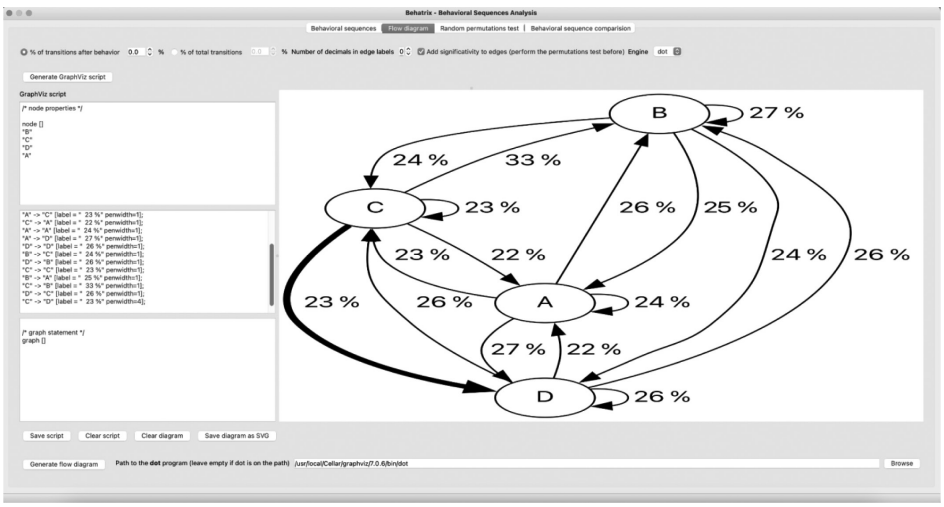


FIGURE 24.5 Flow diagrams in BEHATRIX. Using an observation made in BORIS and subsequently exported with the function 'as behavioural sequences' or copying behavioural strings from a text file or spreadsheet, it is possible to generate a flowchart in BEHATRIX. The flowchart will not only show the behavioural transitions, but can also display the results of a random permutation test and generate thicker connections for those transitions that occur significantly more than others.

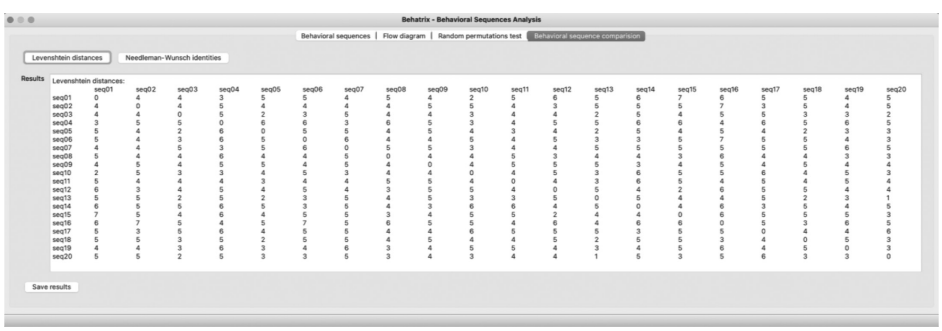


FIGURE 24.6 Levenshtein distance calculation in BEHATRIX. Sequences of behaviour can be subjected to different distance calculations. BEHATRIX allows the Levenshtein distance and Needleman-Wunsch identities to be calculated. In both cases, the output is rendered as a dissimilarity matrix.

by identifying singers and phrase-type combinations in songs. They found that male and female adult indris differed in the phrase organization of their songs and, interestingly, that male contributions were more similar to those of other males than female ones. Moreover, BEHATRIX is useful to compute conditional probabilities of transition between different calls in animal vocalization, creating empirical transition matrices on which can be computed Markov chains. Markov chains (or Markov processes) are stochastic models describing the probability of each event depending on the probability

9780192886491_Book.indb 465 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:04:52

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

of the previous one. This approach allowed researchers to investigate complexity and predictability of indris songs, finding that higher levels of flexibility in the singing of females individuals does not imply higher flexibility in the singing of males individuals, and vice versa (Zanoli et al. 2023)

24.3 Conclusion

Observational studies are a powerful tool for those who wish to formulate non-human and human behaviour hypotheses. Technological advances provide us with increasingly powerful tools capable of accumulating large amounts of data. In the overview we have presented in this chapter, we hope to have clarified the importance of embracing the correct approach concerning the type of results we wish to obtain and to have provided readers with a quick review of some operational aspects that may be useful for those approaching this type of study for the first time. Although there are no shortcuts in the study of behaviour, the quantitative approach using appropriate tools and techniques allows us to maintain a high degree of scientific rigour and guarantees the repeatability of our results.

References

Aychet, J., Blois-Heulin, C., and Lemasson, A. (2021). 'Sequential and network analyses to describe multiple signal use in captive mangabeys', *Animal Behaviour* 182: 203–226.

Bakeman, R. and Quera, V. (2011). *Sequential Analysis and Observational Methods for the Behavioral Sciences*. Cambridge: Cambridge University Press, https://doi.org/10.1017/CBO9781139017343.

Bateson, M. and Martin, P. (2021). *Measuring Behaviour: An Introductory Guide*. Cambridge: Cambridge University Press.

Boersma, P. and Weenink, D. (2020). 'Praat: Doing phonetics by computer [Computer program]', Version 6.0. 37. Retrieved 24 November 2020.

Bolt, L. M. (2020). 'Affiliative contact calls during group travel: Chirp and wail vocalization use in the male ring-tailed lemur (*Lemur catta*)', *Folia Primatologica* 91: 575–594, doi: 10.1159/000508808.

Bousquet, C., Sumpter, A. H., David, J. T., and Manser, M. B. (2011). 'Moving calls: A vocal mechanism underlying quorum decisions in cohesive groups', *Proceedings of the Royal Society B*. 2781482–1488, http://doi.org/10.1098/rspb.2010.1739C4043176.

De Gregorio, C., Zanoli, A., Carugati, F., et al. (2022). 'Parent-offspring turn-taking dynamics influence parents' song structure and elaboration in a singing primate', *Frontiers in Ecology and Evolution* 10: 1–14.

De Gregorio, C., Carugati, F., Valente, D., et al. (2022). 'Notes on a tree: Reframing the relevance of primate choruses, duets, and solo songs', *Ethology Ecology & Evolution* 34(3): 205–219.

Friard, O. and Gamba, M. (2016). 'BORIS: A free, versatile open-source event-logging software for video/audio coding and live observations', *Methods in Ecology and Evolution* 7(11): 1325–1330.

9780192886491_Book.indb 466 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:04:52

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Friard, O. et al. (2023). 'Behatrix: Behavioral sequences analysis with permutations test', http://www.boris.unito.it/pages/behatrix, accessed 4 February 2023.

Gooskens, C. and Heeringa, W. (2004). 'Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data', *Language Variation and Change* 16(3): 189–207.

Heesen, R., Hobaiter, C., Ferrer-i-Cancho, R., and Semple, S. (2019). 'Linguistic laws in chimpanzee gestural communication', *Proceedings of the Royal Society B* 286(1896): 20182900.

Jäger, G. (2015). 'Support for linguistic macrofamilies from weighted sequence alignment', *Proceedings of the National Academy of Sciences* 112(41): 12752–12757.

King. S. L., Guarino, E., Donegan, K., et al. (2021). 'Evidence that bottlenose dolphins can communicate with vocal signals to solve a cooperative task', *Royal Society Open Science* 8: 202073, https://doi.org/10.1098/rsos.202073.

Levinson, S. C. and Holler, J. (2014). 'The origin of human multi-modal communication', *Philosophical Transactions of the Royal Society B: Biological Sciences* 369(1651): 20130302.

Mann, R. P., Faria, J., Sumpter, D. J., and Krause, J. (2013). 'The dynamics of audience applause', *Journal of The Royal Society Interface* 10(85): 20130466, doi: 10.1098/rsif.2013.0466.

McGuire, J. F., Sturm, A., Ricketts, E. J., et al. (2022). 'Cognitive control processes in behavior therapy for youth with Tourette's disorder', *Journal of Child Psychology and Psychiatry* 63(3): 296–304, doi: 10.1111/jcpp.13470.

Price, T., Wadewitz, P., Cheney, D., et al. (2015). 'Vervets revisited: A quantitative analysis of alarm call structure and context specificity', *Scientific Reports* 5: 13220, https://doi.org/10.1038/srep13220.

Rasouli, A., Kotseruba, I., and Tsotsos, J. K. (2017). 'Agreeing to cross: How drivers and pedestrians communicate', in *2017 IEEE Intelligent Vehicles Symposium (IV)* 264–269, IEEE.

Scheun, J., Bennett, N. C., Ganswindt, A., et al. (2015) 'The hustle and bustle of city life: Monitoring the effects of urbanisation in the African lesser bushbaby', *The Science of Nature* 102: 57, https://doi.org/10.1007/s00114-015-1305-4.

Seyfarth, R. M., Cheney, D. L., and Marler, P. (1980). 'Vervet monkey alarm calls: Semantic communication in a free-ranging primate', *Animal Behaviour* 28: 1070–1094.

Spezie, G., Torti, V., Bonadonna, G., et al. (2023). 'Evidence for acoustic discrimination in lemurs: A playback study on wild indris (*Indri indri*)', *Current Zoology* 69(1): 41–49, https://doi.org/10.1093/cz/zoac009.

Valente, D., De Gregorio, C., Favaro, L., et al. (2021). 'Linguistic laws of brevity: Conformity in *Indri indri*', *Animal Cognition* 24: 897–906.

Valente, D., Torti, V., De Gregorio, C., et al. (2023). 'First evidence of contagious yawning in a wild lemur', *Behavioral Ecology and Sociobiology* 77(1): 11.

Zanoli, A., De Gregorio, C., Valente, D., et al. (2020). 'Sexually dimorphic phrase organization in the song of the indris (*Indri indri*)', *American Journal of Primatology* 82(6): e23132.

Zanoli, A., Raimondi, T., De Gregorio, C., et al. (2023). '"The song remains the same": not really! Vocal flexibility in the song of the indris', *Animal Cognition*, https://doi.org/10.1007/s10071-023-01826-6.

9780192886491_Book.indb 467 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:04:52

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

9780192886491_Book.indb 468 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:04:52

CHAPTER 25

RE-IMAGINING GREAT APE GESTURE (CODING)

ERICA A. CARTMILL, EMILIE GENTY, KIRSTY E. GRAHAM, CHARLOTTE GRUND, AND CATHERINE HOBAITER

25.1 Introduction

Great ape gesture has been a source of scientific curiosity for over 150 years (Cartmill and Hobaiter 2019a). Darwin himself wrote detailed descriptions of ape gestures, describing them as more 'expressive' than facial expressions, foreshadowing contemporary work on their flexibility and semiotic features (Darwin 1872; Plooij 1977; Tomasello et al. 1994; Cartmill and Byrne 2010; Hobaiter et al. 2022). The past forty years have brought an explosion of interest, initiated in part by the recognition that ape gesturing shares fundamental commonalities with human language production. This connection was crucial in developing scholarship linking ape gesture and the evolution of language, focusing research efforts on cognitive-linguistic questions of intentional use, acquisition, and meaning (Tomasello et al. 1985, 1994; Pika et al. 2003, 2005; Liebal et al. 2006; Genty et al. 2009; Hobaiter and Byrne 2011, 2014; Roberts et al. 2014; Frohlich et al. 2016a; Byrne et al. 2016; Graham et al. 2018, 2020; Bard et al. 2019).

The surge in gesture studies brought with it a range of diverse new methods and approaches. For example, studies vary substantially in the actions they group as a single category or unit, for example: different kinds of 'touches' can be considered a single type of gesture (e.g. Tanner and Byrne 1996) or many different gesture types distinguished by body parts and context (e.g. Bard et al. 2019 distinguish thirty-six forms of touch). Methodological breadth, including different techniques, species, populations, and settings, allows multiple perspectives and enhances both the robustness of our conclusions and their generalizability. But such diversity can also come at

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

a cost, particularly where sources of methodological variation are opaque. Rodrigues and colleagues (2021) found that almost a third of gesture studies did not provide basic definitions of which actions constituted a gesture, making replication and comparison highly challenging. While some areas of gesture research diversified, other recent research efforts have become less diverse, moving away from much of the flexibility and expressiveness that made ape gesture notable to early scholars. For example, features that might contribute to what Darwin referred to as *expressiveness*—like the relative size or salience of signals—are rarely remarked upon in current work, unless perhaps as a feature that apes manipulate in modifying signals to be more easily perceived by others (e.g. Cartmill 2009, 2023; Badihi et al. 2023).

25.1.1 The case for going bottom-up

When addressing large-scale comparative questions, particularly those that trace the evolution of specific characteristics of communication systems, the ability to compare like-with-like becomes crucial (Rodrigues et al. 2021). Many species that gesture—whether apes or elephants—are long-lived, highly social, and behaviourally diverse. It can be difficult to track individuals across time or match individuals across datasets, and so comparative questions about gesture are typically addressed by comparing across studies and research groups. One significant advantage of contemporary gestural research is that—collectively—we now have the datasets needed to re-address gestural questions at scale. Large-scale data allows comparison across populations, groups, demographics, individuals, and contexts. It also allows researchers to deepen their understanding of generalized features of the ape gestural system. Excitingly, it also offers the opportunity to address questions about sources of phylogenetic variation in gesture form and use, and to trace use across individual life histories, building a roadmap on which we can identify both evolutionary and proximate factors shaping gesture use.

Combining datasets from across studies provides an effective means to increase and diversify our sample, but aligning definitions, methods, and formatting across those studies presents a significant hurdle. One solution to increasing alignment between datasets would be for different researchers to adopt the same methods. But—even if this were plausible—this restriction would likely come at a substantial scientific cost to diversity and perspective. The opportunity to compare, contrast, and innovate is essential to progress in research. Another approach is to re-imagine the way in which we code gesture. Our tendency—following other studies of behaviour—has been to construct categories from the top down. This approach requires researchers to decide what behaviours are of interest and pre-assign our observations to particular 'types'. One potentially transformative approach to describing gesture would be to do the opposite: to construct categories from the bottom up. While there have always been 'lumpers' and 'splitters' in gesture coding, adopting a bottom-up approach is about more than the level of granularity at which we define gesture types. Highly split repertoires, when defined from the top down, are often still vulnerable to inconsistencies in how splitting rules

9780192886491_Book.indb 470 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:04:52

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

are applied. For example, in one of our own papers, chimpanzee gestures were initially classified into 115 types and later lumped to sixty-six (Hobaiter and Byrne 2011). However, within both levels of granularity some, but not all, gesture actions were split by context (e.g. Present grooming, Present sexual); and some, but not all, gesture actions were split by their use of repetition (e.g. Tap, Tapping), and so on. Building from the bottom up allows researchers to a) systematically (re-)combine features of a gestural system into gestural units at different levels of granularity, and b) tailor gestural units to specific research questions in a transparent way without needing to recode original observations.

Similarly, bottom-up coding allows us the flexibility to construct elements of gestural datasets that are relevant to specific theoretical concepts, and to adapt them across different studies and over time as our understanding of these concepts evolves. Central concepts applied to gesturing such as *intentional use* have been refined over the years as our understanding of apes' cognitive capacities developed (e.g. Theory of Mind: Premack and Woodruff 1978; Heyes 1998; Call and Tomasello 2011; Krupenye and Call 2019; see also Leavens et al. 1996, 2005, Tomasello et al. 1997). The use of intentionality criteria was essential to discriminate gesture from non-communicative movements that are similar in form. For example, there is nothing inherently different in form between raising your arm to reach something above you and producing an Arm Raise gesture. But they can be discriminated by considering accompanying behaviour, such as gaze or response-waiting, that provides evidence of an intent to communicate. The general methods for coding intentionality have changed relatively little since their adaptation from the study of linguistic development in human infants (e.g. Bates et al. 1975). Recently, application of these intentionality criteria to non-primate species has revealed seemingly intentional gesturing in fish (Vail et al. 2013) and highlighted intentional use of some ape vocal signals (e.g. Schel et al. 2013), challenging prior findings that primate vocalizations were largely inflexible. These studies have raised questions about whether intentional signalling is widespread across species and communicative channels, or whether our methods do not allow us to make relevant distinctions between audience-sensitive use and more language-like intentional use (Townsend et al. 2017). Bottom-up coding of the behaviour accompanying gesture—for example annotation of interactional features such as the type and timing of gaze and pauses—allows us to introduce new measures of intentionality as our criteria for intentional use evolve (again without necessitating recoding of our data).

While we feel there are clear benefits, it is important to recognize that bottom-up coding comes with costs, substantial ones in the case of coding effort and time, but also potentially in terms of intercoder reliability, where increasingly fine-grained or subtle distinctions may require additional training or substantial experience to reliably discriminate. However, one advantage to this approach is that variables that show variation across coders at a highly split level, may be more robust to variation at a more lumped level. If interobserver reliability is assessed at both levels, the outcome can guide decisions about an appropriate level for analysis—without the need for recoding of data—and can clarify where further attention is needed to the coding scheme or definitions in future work.

9780192886491_Book.indb 471 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:04:52

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

25.1.2 Gesture *in situ*—studying gesturing is about more than gesture form

Our ability to describe not only the physical forms of gesture units, but the way in which they are used—gesturing—requires that we consider gesture *in situ*. In ways, this harkens back to the early descriptions of ape gesturing, which highlighted their flexibility and expressiveness. Most studies of gesture already incorporate the simpler and more stable features of the individuals involved in gestural interactions, like sex, age, and perhaps kinship of signaller and recipient (e.g. Tomasello et al. 1994; Frohlich et al. 2016b; Hobaiter et al. 2017). A richer contextualization of the nuanced information in gesturing requires that we start to attend to additional extragestural elements of interaction, some of which are stable (sex and kinship), some of which are changeable but predictable (age, social relationships, ontogenetic experience, reproductive status, physical environment), and some of which are highly dynamic (audience, arousal, behavioural context, immediate past experiences; Graham et al. 2022).

Most studies on ape communicative systems have focused on one communicative channel in isolation: gestures, vocalizations, or facial expressions (Zuberbühler 2005; Byrne et al. 2016; Waller et al. 2017). However, gestures—like most species' signals—do not occur in isolation, but as part of an interwoven tapestry of other signals, that include calls, facial expressions, olfactory cues, and more (e.g. Partan and Marler 2005; Higham and Hebets 2013; Liebal et al. 2014). Studies of vocalization often already include more dynamic features of the interactants such as relative social relationship, rank, reproductive status, or audience composition (e.g. Slocombe et al. 2007; Crockford et al. 2012; Fedurek et al. 2015). If gesture research incorporated these features as well, it would not only expand the types of possible analyses of gesture but may also improve the ability to directly compare gesture with other channels. In the simplest extension, this would involve coding non-gestural signals with established ethograms, like vocalizations and facial expressions (e.g. Hobaiter et al. 2017; Wilke et al. 2017). In others, this would necessitate creating new measures to include things like olfactory or gustatory cues that our current methods do not capture. It would also require thinking about the ways in which communicative acts are embedded within physical spaces and laminate within ongoing interactions (e.g. Goodwin 2018).

25.2 Key innovations in gesture coding—and how they shape the methods we use to explore gesture

Four years ago, frustrated by what we felt was a stagnation in where the study of ape gesture was going, we wrote a call to action:

9780192886491_Book.indb 472 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:04:52

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

> *We urge scholars to collect new data and leverage existing data in ways that maximize the potential for comparison across datasets and articulation with studies of other communicative modalities … we advocate for a transition away from using intentionality as a marker of the 'special status' of gesture, and towards using gesture as a window onto the lives and minds of apes.* (Cartmill and Hobaiter 2019b)

This was as much about holding our own work to account, as it was about encouraging others. We have developed this call to action across three themes below: section 25.2.1, Repertoires and gesture 'types'; section 25.2.2, Multichannel communication; and section 25.2.3, Intentionality, meaning, and moving beyond it.

25.2.1 Repertoires and gesture 'types'

In bottom-up coding, we aim to code gesture at a fine-grained level, breaking concepts (such as intentional use) and categories (such as gesture types), down into their component parts. This approach relies on coding the behaviour we observe, rather than direct coding of the cognitive categories that we infer using these behavioural cues. For example, observable actions such as 'pause in communication' and 'visual monitoring of recipient' are behavioural cues that we use to infer the state of 'response waiting'. We should be clear that strictly bottom-up coding is essentially impossible to attain; describing any behaviour requires us to impose some sort of *a priori* categories. Importantly, many of the features we are attempting to capture exist on a graded continuum: different actions grade into one another (a back-and-forth Swing into a back-and-forth Shake); size, speed, vigour, and duration are all continuous, and there are inherent limitations to defining and contrasting gestures and gesture features as 'units' rather than as continuous variables. While they can be split into increasingly fine-grained categories, almost all professedly bottom-up coding schemes (including our own, e.g. Grund et al. 2023) categorize some features in advance, giving human coders a starting point for identifying movements worth coding. Whenever we break a 'set' or category down into its component parts, we add to the already substantial investment of labour needed to code data. But, where that investment is practical, substantial long-term benefits accrue.

For one, bottom-up coding avoids or minimizes the tedious recoding necessitated when new understandings demand fundamental reconsideration of the data. For example, where a gesture 'type', such as a *Reach* needs to be split to test a new hypothesis (e.g. whether meaning within Reaches is discriminated by flexion at the wrist). Coding the different parts of the gesture's physical form separately—body part, movement, flexion, orientation, directionality, etc.—from the start appears uneconomical but, in the long run, allows researchers to reconstitute their data, without recoding, at the level needed to test novel questions. The same basic principle holds across many of the aspects of gesturing and the gestural system.

The imposition of pre-defined categories on observations also introduces anthropocentric (and other species-centric) biases into our analyses. It is tempting to think that this

9780192886491_Book.indb 473 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:04:52

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

is not as much of a problem for researchers of ape behaviour as it is for people studying crows, dolphins, or squid. Because apes share our basic body plan—they are, in the literal sense of the word, anthropomorphic—we can imagine in some detail what it is like to be an ape. We know what movements are physically possible for them, and we tend to assume that we know which ones are relevant to signal production and which represent incidental variation with little-or-no semiotic content. But this belief may seduce us into misunderstandings that students of cetaceans or cephalopods have no trouble avoiding. In coding from the bottom up, we can minimize these *a priori* human-centric biases from the very beginning of our study of a new system or species (Grund et al. 2023). In an early example of this, we re-examined our assumptions about what distinguished different 'types' of chimpanzee gesture by breaking down gestures into their component parts and exploring whether clusters of gesture form within a type mapped onto different patterns of use (Hobaiter and Byrne 2017). In some cases, our original top-down approach based on anthropocentric assumptions—for example, distinguishing rhythmic properties such as synchronous hitting from alternating hitting, or specifying whether a particular action is performed with the foot or hand—turned out to be features of gestural behaviour apparently irrelevant to the chimpanzees using them.

Similarly, much of our understanding of ape gesturing comes from the in-depth study of chimpanzees (Rodrigues et al. 2021); however, we have increasing evidence that the gestural system is expressed differently in different species of apes. Arboreal orangutans produce the same gestural movements with a greater range of limb movements than the more terrestrial chimpanzees (Knox et al. 2019). Our current assumptions about what distinguishes one gesture unit from another, for example, whether changing limb represents the variable expression of a single unit, or distinguishes a new unit, may depend in part on the affordances of the socioecological environment to which a species is adapted. Similarly, deciding whether additional gesture units represent further gestures within a communicative act, as opposed to the initiation of a new communicative act, has typically been done by identifying 'appropriate' timings for response-waiting. With a range of studies suggesting species differences in responsiveness (bonobos and chimpanzees: Fröhlich et al. 2016; mountain gorillas and chimpanzees: Grund et al. 2023; orangutans: Knox et al. 2019), it is important to avoid structural biases in coding not only towards expectations derived from human experience and embodiment, but also towards expectations derived from the intensive study of *Pan troglodytes*.

25.2.1.1 No one repertoire to rule them all

Within repertoires, tactile gesture forms present special challenges to the coder. While we can observe what body part the signaller uses to make contact, which body part of the recipient is contacted, and for how long, many features of tactile gestures are continuous variables and largely inaccessible to researchers (e.g. what force is applied, what quality of touch, and how the tactile stimulus varies across the contact). These types of signals are also not well suited to our current intentionality criteria (strongly biased towards visual attention; Dafreville et al. 2021) and are challenging to sort into categories based only on the visual and auditory information contained in a video recording.

9780192886491_Book.indb 474 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:04:52

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

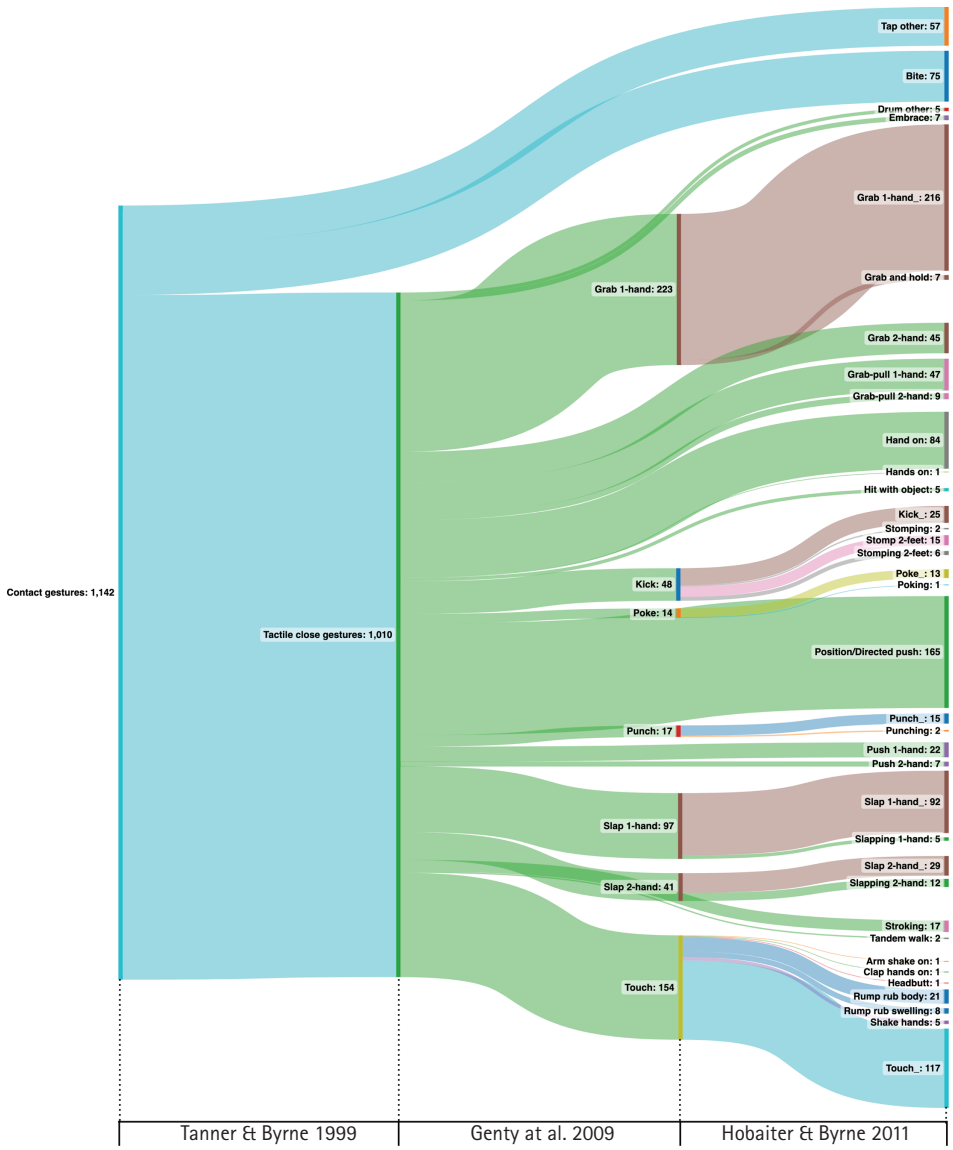


FIGURE 25.1 Touching on contact gestures.

Here we use data from a set of n = 1,142 instances of chimpanzee contact gestures and depict increasing granularity over time, following the coding schemes of Tanner and Byrne (1996) (n = 3 contact categories); Genty et al. (2009) (n = 23 categories); and Hobaiter and Byrne (2011) (n = 37 categories). If the same gestures were split according to our latest coding scheme (Grund et al. 2023) the categories (or 'morphs') could be constructed from twenty-two different types of tactile gesture action that are produced by or make contact with twenty-six different body parts, with additional options for further discrimination by modifiers such as repetition, etc., allowing even greater detail.

9780192886491_Book.indb 475 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:04:53

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

The signaller's intended meaning when producing a tactile gesture, or the recipient's interpretation of it, might also depend on the identity of the signaller–recipient pair. What might be a very hard slap when delivered by an adult gorilla to a juvenile, might be experienced as a much softer one by an adult recipient, or might be interpreted differently based on the recipient's social rank or relationship to the signaller (Heesen et al. 2021; Rodrigues et al. 2022). Early coders addressed this uncertainty by applying a very high level of lumping; for example, using a single 'tactile close' category of gestures (Tanner and Byrne 1996). As more data has accumulated, more specific types of contact action have been parsed out (Figure 25.1), splitting gesture types and increasing tactile repertoire sizes across studies.

At the end of the day, there is no 'correct' level of granularity. The choice of category sizes and grouping for data analysis needs to reflect the research question and data availability (Cartmill and Byrne 2011). While fine-grained coding is ideal, the most critical aspect is to apply a consistent level of granularity. For example, in earlier work (Hobaiter and Byrne 2011) some gesture actions (Grab) were split by body part (one-hand, two-hands) while others (Punch) were not. A feature of our latest coding framework (Grund et al. 2023) is that each gesture action is systematically split by modifiers such as body-part signaller, body-part contact, or object use, allowing more flexibility in recombination and comparison, and a more consistent use of splitting higher-order gesture 'actions' into what we term gesture 'morphs'. There are no rigid rules about whether and when a modifier creates variation that might be described as a new unit (for example, variation in dialect) or variation akin to expression (as in variation in accent or emphasis). The specific effect of a modifier could vary, most likely by species, but also potentially by social unit, context of information, or even between different types of gesture actions. We recommend starting out by considering what the desirable set of modifiers and the resultant morphs should be, given the research question, and then (in all likelihood) lumping up from there in a trade-off given the necessity of a particular detail to address the question of interest, and the data and coding time available. No matter what approach is adopted, it is crucial that concepts and methodological choices are made transparent. If they are not, meaningfully comparing gestural repertoires becomes practically impossible.

25.2.2 Multichannel communication

The benefits of studying how apes combine their different signal channels (gesture, vocalization, facial expression, and more) is increasingly acknowledged in the scientific community (e.g. Slocombe et al. 2011; Higham and Hebets 2013; Liebal et al. 2014; Fröhlich and van Schaik 2018), with an increasing number of studies focusing on multimodal and multichannel signal use (e.g. Leavens et al. 2010; Genty et al. 2014, 2015; Levinson and Holler 2014; Taglialatela et al. 2015; Wilke et al. 2017; Hobaiter et al. 2017; Genty 2019). A number of recent studies have compared combinations of signals across channels to signals produced in a single channel, including descriptions of (a) frequency

9780192886491_Book.indb 476 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:04:53

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

(Leavens and Hopkins 2005; Pollick et al. 2008; Leavens et al. 2010; Genty et al. 2014; Wilke et al. 2017); (b) efficiency (Genty et al. 2014, 2015b; Wilke et al. 2017); (c) social function (Genty et al. 2014, 2015b; Wilke et al. 2017; Hobaiter et al. 2017); and d) functional specificity (Genty et al. 2014, 2015b; Genty 2019).

Research into multichannel communication remains in its infancy, and we have much to learn about when and why apes combine signals. Given the size and flexibility of apes' communicative repertoires (e.g. Clay et al. 2015; Byrne et al. 2016), more than 1,000 multichannel combinations are theoretically possible, but only a few are used regularly (see Wilke et al. 2017). It might be the case that gestures are combined with signals in other channels to increase specificity by refining the message conveyed (Genty et al. 2014, 2015b; Genty 2019). But we do not know why some combinations are used and others are not. Multichannel signal combinations are typically rare, but appear more likely to elicit responses as compared to single-channel vocal signals (Leavens and Hopkins 2005; Leavens et al. 2010; Genty et al. 2014, 2015b; Wilke et al. 2017; Hobaiter et al. 2017). In contrast, single-channel gestural signals are similarly likely to receive a response as when combined with another signal channel (Wilke et al. 2017). This pattern resembles that of human language use, in which listeners tend to rely more on co-speech gestures when they communicate different information than the accompanying speech (Thompson and Massaro 1986; Goldin-Meadow et al. 1992; Cassell et al. 1999).

One significant hurdle to a deeper understanding of signal combinations in other apes is that the different sensory channels have typically been studied using substantially different methods. Vocalizations and facial expressions can be classified with relatively straightforward descriptions of the physical form of the signal; however, gesturing requires that a particular action be accompanied by other behaviour to discriminate a movement used as a communicative signal, from one used for effective action on the world or other individuals in it. As a result, while studies of vocalization can consider signallers in isolation, gesturing necessitates the study of the interaction between signaller and recipient. A knock-on effect of these structural differences in approach is that even relatively fundamental aspects, such as what indicates the start and end of a communication interaction, tend to vary between the study of different channels of communication. Consider a chimpanzee feeding in a tree, producing a near-constant stream of food-grunts. These vocal signals may vary in intensity and acoustic structure but are ongoing over a twenty minute period. Within this event, after forty-five seconds, another chimpanzee approaches and is rebuffed with a Hand-fling gesture ('move away'), fifteen minutes later, but within the same stream of vocalizations, an ally approaches and is given a gentle touch on the shoulder as they walk by. Gestural researchers use pauses in communication, or changes in activity to parse out communicative events—and would certainly want to separate the two gestural events (different recipients, different meanings). But when the vocal stream is added, the lines are blurred—the two sets of gestures are now connected. As we consider more and more information in an interaction, our ability to define and classify particular events becomes harder. Again, one way to

9780192886491_Book.indb 477 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:04:53

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

address this challenge is to code in such a way that the events can be classified (and reclassified) at a later stage—and 'in-video coding' tools provide us with a powerful means to do so.

Software such as Observer XT, ELAN, Solomon coder, or BORIS link behavioural coding to the associated video files, with each coded variable time linked to a moment in a video file, benefiting both multichannel analysis and bottom-up coding methods. In-video coding allows for accurate and fine-grained annotations of all signals and related behaviours, also facilitating reproducibility and intercoder reliability tests, and allowing researchers to make their data more accessible to others. Our most recent coding schemes have been implemented in ELAN: a user-friendly, free, open-source software package that was designed by the Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands, to facilitate the study of spoken and signed languages, and gestures (ELAN (Version 6.4), 2022; https://archive.mpi.nl/tla/elan). Particularly

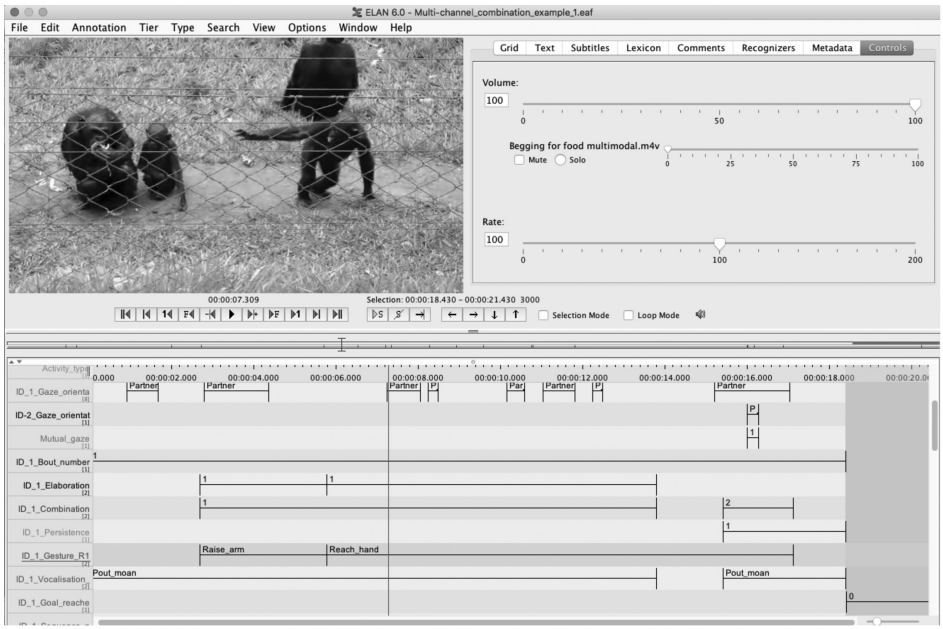


FIGURE 25.2 Example of ELAN coding of a bonobo interaction involving a multichannel combination (vocalization + gesture) to request food.

In this clip, a juvenile male bonobo (BLG) is standing quadrupedally at about 1.5 m from an unrelated adult female bonobo (SLG) who is feeding. SLGG is not visually attending to BLG. BLG initiates the interaction with SLG by gazing at her. BLG produces a pout moan (vocalization, audible). After 690 ms, BLG adds an Arm Raise (gesture, silent-visible) to the pout moan, immediately followed by an Arm Reach (gesture, silent-visible). The addition of the Arm Raise and Arm Reach gestures constitutes both a communicative elaboration and a multichannel combination. BLG holds his arm extended to SLG for a total of 11.35 sec. During this hold, he pauses his vocalization for 1.6 sec while still holding his arm out, then starts pout moaning again. This repetition of the pout moan vocalization constitutes communicative persistence, and combined with Arm Reach, would be considered a second multichannel combination in the communicative bout. After a total of 18.4 sec, BLG stops signalling and sits down. SLG ignores him. BLG's goal (food sharing) was not met, however he does not persist in communicating further.

Credit: Emilie Genty.

9780192886491_Book.indb 478 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:04:53

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

well suited for the study of great ape gestures, and multichannel signalling in particular, ELAN users create 'tiers' for each variable and the timings and overlap between annotations are automatically captured. These built-in links from code to video allow for quick and easy comparison of different variables, by forming smaller more highly tailored corpora. Importantly, ELAN users can generate reusable and sharable coding templates, that can be duplicated or adapted depending on the research question (e.g. GApS coding scheme, GestureOrigins coding scheme; Genty et al. 2023, Grund et al. 2023). ELAN can also be integrated with other open-source coding tools. For example, there is a built-in option to incorporate 'Praat' for acoustic analyses (Boersma and Weenink 2022). The ELAN community has also developed tools to facilitate the management of ELAN annotated files. Elan_finder is a free open source software providing a rapid way to filter, retrieve, preview, and export ELAN annotated files from large databases (https://github.com/MitchFuchs/elan_finder.git). Our published coding frameworks give a much more detailed breakdown of how to get stuck into detailed coding, but below we outline a brief example of a multichannel communicative event in the context of feeding (Figure 25.2) to highlight how different features of an interaction are captured.

25.2.3 Intentionality, meaning, and moving beyond it

Many of the methods we use to identify cognitive capacities in human language rely on the use of language as a research tool. As it is not possible to ask other apes about their communicative acts directly, non-verbal methods must be used to infer cognitive states. To this end, the methods originally developed for studying pre-verbal human children have been productively applied to the study of apes, with cross-fertilization from infant research particularly fruitful in studying theory of mind, sensitivity to the visual attention of others, and intentionality (e.g. Bates et al. 1975; Premack and Woodruff 1978; Tomasello et al. 1985, 1997; Call and Tomasello 2011).

It is important to clarify here that we do not use *intentional* to necessarily mean 'under voluntary control', rather intentional communication is communication that indicates the signaller's intention. Intentionality can be understood as *aboutness* or *goal-directedness*. With it comes much of what we recognize in the use of language, including the ability to mean different things by using the same signal in different ways and in different contexts. Intentionality has been a critical topic since researchers first began to study ape gestures. Identifying which movements constitute gestures relies largely on determining which movements are used to communicate. It is possible for researchers studying vocalization to remain largely agnostic to intention as whether an animal is vocalizing is purely a question of whether the vocal cords are engaged. Determining whether a wave of the arm is a gesture or a functional arm stretch depends on whether the movement is used to communicate, and this question is entangled with the signaller's intention.

9780192886491_Book.indb 479 /12_first_proofs/first_pr…/xml_for_typesetting 27-Mar-25 23:04:53

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

When the intentionality of ape gestures started to be systematically studied in the 1980s (e.g. Tomasello et al. 1985, 1994) it represented a radically new claim about animal communication, one that still comes under appropriate scrutiny (e.g. Scott-Phillips 2016). Researchers have worked to discriminate the evidence for zero-order intentions (signaller broadcasts their signal), first-order intentions (signaller's goal is to change a specific recipient's *behaviour*), and second-order intentions (signaller's goal is to change a specific recipient's *mind*) (Grice 1989; Cartmill and Byrne 2010; Hobaiter and Byrne 2014; Hobaiter et al. 2022). However, current methods to identify intentionality in communication, including gesture, remain fundamentally the same as those used in the 1970s. We look for patterns of gaze and pauses in behaviour that can be used to infer attention to a particular recipient and their behaviour, and often include waiting for a response (Bates et al. 1975). Inferring intentionality relies on multiple instances of a particular gesture. The consistent pattern of use across cases, the flexibility that the signaller shows in selecting the appropriate modality of communication for a particular recipient, the regular and repeated use of pauses, together indicate the presence of intentional communication. Once researchers identified which movements were used in goal-directed ways to communicate with others, they were able to ask what the apes were communicating about. If gestures are indeed intentional, we can go beyond questions of what informational content they carry and ask, in the linguistic sense, what does the signaller 'mean' to communicate when using them.

Previous studies of non-human communicative signals, including ape gesture, tended to categorize signal information on the basis of the context in which it was produced. This is a *feeding* call, a *travelling* gesture, a *play* facial expression. Here, though, there is a substantial challenge for addressing questions of meaning. The signallers in question may be communicating in different contexts (feeding, travelling, playing) but could all have 'meant' the same thing, e.g. 'come here'. The context alone does not reveal the signaller's goal within that context. One could also equate meaning or function with the recipient's immediate response (e.g. Roberts et al. 2014), but these responses could include refusal, misunderstanding, or even negotiation and compromise. Instead, the most effective method available is to describe the change in the recipient's behaviour that stops the signaller from signalling. Now referred to as the Apparently Satisfactory Outcome (ASO), this method was first applied to explore orangutan (Cartmill and Byrne 2010) and then gorilla (Genty et al. 2009) gesture in captive groups, before being applied to wild apes (Hobaiter and Byrne 2014; Graham et al. 2018).

As an example, imagine that Alice would like Fred to pass her the cake at dinner. Alice asks Fred to pass it to her, and then waits for his response. She would be unlikely to simply repeat 'cake, cake, cake'. If Fred didn't respond, or if he passed Alice something else—the cheese, let's say—Alice would repeat her request, perhaps adjusting it to elaborate or clarify (rephrasing, or adding a pointing gesture). In this interaction we see the characteristics of intentional use—a signal selected to be detectable, a pause while the signaller waits for a response from their recipient (they are communicating *to someone*), adjustment or elaboration of the communication where it failed—but the one thing that is likely to make Alice stop asking for the cake is when Fred passes it. Of note, Alice

9780192886491_Book.indb 480 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:04:53

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

is likely to stop communicating not when she eats the cake (i.e. when the motivation that initiated the signal is more directly addressed) but when she realizes that Fred *understands* her signals.

Together—the recognition of intentional use, and the description of gestural meanings—have provided some of the key insights into the similarities and distinctions between ape gesture and human language. We have evidence for intentionality and gesture meaning in every individual, population, and species studied so far. But it may be time to move beyond these measures.

Having established that non-human apes use gestures intentionally, limiting our research questions only to gesture cases that pass strict intentionality criteria may severely restrict our understanding of gesturing (Cartmill 2023). There are undoubtedly important nuances that remain to be decoded. What do we do with the signals that look like gestures, but that are not accompanied by easily observable markers of intentional use, such as stereotyped displays or long-distance communication (e.g. chimpanzee tree-buttress drumming)? Gestures and other signals that are centred on audible or tactile modalities do not necessarily need to be produced with sensitivity to the recipient's visual attention (a key criterion for intentional use; Dafreville et al. 2021). These signals are often overlooked or excluded from current analyses, but there is likely a continuum of intentional signal production depending on support from extra communicative cues.

Intentionality in ape gesturing was once a relatively controversial claim that needed substantial supporting evidence. However, there is now such strong evidence for apes' capacity for (at least first-order) intentional gesturing, that perhaps we can start to examine potentially communicative gestural behaviours that do not meet the strict criteria for intentional gesturing. Doing so would free up important new avenues of exploration. Current analyses of meaning are largely limited to imperative requests because they rely on an explicit behavioural response as an indication of what the signaller wanted to achieve. If we were able to assume that a gesture is produced intentionally, but we see no apparent response and no subsequent persistence and elaboration, we might consider whether it carries non-imperative meaning (e.g. pragmatic or (proto)declarative; Hobaiter et al. 2022, Graham et al. 2022).

Here, again, bottom-up coding methods allow us the opportunity to apply our current methods for discriminating meaning and to expand them to encompass additional questions. Coding observable behaviours about all parties involved in an interaction allows researchers to (re)construct criteria for defining social interactions and gesture meaning in flexible ways. It could include aspects of context before and after the communicative act, the identity, age, sex, and relationship between signaller and recipient, as well as information about the audience. Where longitudinal data are available, it may be fruitful to include information about previous interactions and ask how past interactions shape future ones (e.g. Wittig and Boesch 2003). This flexibility can be seen in a recent study of bonobo gestures in which incorporating the behavioural context of the signaller before the gesture, as well as the age/sex of signaller and recipient, proved important in eliminating previously described ambiguity in signaller meaning (Graham et al. 2020).

9780192886491_Book.indb 481 /12_first_proofs/first_pr[illegible]/xml_for_typesetting 27-Mar-25 23:04:54

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Crucially—in taking a step back from focusing on intentional use—researchers would be freer to integrate aspects of emotion and affect that are commonly used as reasons to exclude interactions from analysis and explore the impressive flexibility and expressiveness of gesture that first captured the imaginations of scientists over 150 years ago.

25.3 Summary and next steps

Similarities between the intentionality and flexibility of ape gestural communication and human language fuelled debates about a possible gestural origin of language. Early methodological approaches focused on providing systematic evidence for the presence of core linguistic concepts—such as intentionality and means-end dissociation—characteristics that had been difficult to establish in primate vocalizations and facial expressions. In consequence, the study of ape gestures, vocalizations, and facial expressions were siloed and employed very different methods. Recent work on multichannel signalling, aided by in-video coding software, has attempted to break down these barriers, resulting in greater collaboration and methodological cross-fertilization. Bottom-up coding further facilitates the consolidation of these previously siloed areas of research.

To address a wider range of questions about ape gesture—including multichannel use, signal combinations, different markers of intentional use, and socioecological context, to name but a few—we benefit from bottom-up coding approaches that minimize *a priori* structural choices and (anthropocentric or pancentric) conceptualizations about data categories. Bottom-up coding is more labour intensive than searching for previously defined categories, but new machine learning tools will allow us to locate relevant video (species and individual detection), identify sections with social interaction (behavioural classification), and perhaps even code some aspects of movement and timing (pose-estimation) in the near future (Wiltshire et al. 2023; Fuchs et al. 2023). These tools are not (yet) driving the research questions, but could instead ease the burden of intensive coding in ways that would give researchers the flexibility to address a wider range of questions, for example detecting multidimensional patterns of variation in movement (Mielke et al. 2024).

Exploring ape communication in open, data-driven, and transparent ways can help us to discover new aspects of ape gesture. Building in collaborative procedures that facilitate data comparison and sharing will allow us to study the gestural communication of these long-lived, socially flexible, and behaviourally diverse species in ways that will dramatically move the field forward and enrich our understanding of the role of gesture in their, and our, communication.

Acknowledgements

The authors would like to acknowledge their many collaborators and colleagues in the field of gesture and beyond for the productive discussion and debates that have led to these ideas, as

9780192886491_Book.indb 482 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:04:54

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

well as the apes we work with for being a constant source of inspiration for the need to do better. CH, KEG, CG received funding from the European Union's 8th Framework Programme, Horizon 2020, under grant agreement no 802719.

References

Badihi, G., Graham, K. E., Fallon, B., et al. (2023). 'Dialects in leaf-clipping and other leaf-modifying gestures between neighbouring communities of East African chimpanzees', *Scientific Reports* 13(1): 147.

Bard, K. A., Maguire-Herring, V., Tomonaga, M., and Matsuzawa, T. (2019). 'The gesture "Touch": Does meaning-making develop in chimpanzees' use of a very flexible gesture?', *Animal Cognition* 22: 535–550.

Bates, E., Camaioni, L., and Volterra, V. (1975). 'The acquisition of performatives prior to speech', *Merrill-Palmer Quarterly of Behavior and Development* 21(3): 205–226.

Boersma, P. and Weenink, D. (2022). 'Praat: A system for doing phonetics by computer, 2000', Software available at www.praat.org.

Byrne, R. W., Cartmill, E., Genty, E., et al. (2016). 'Great ape gestures: Intentional communication with a rich set of innate signals', *Animal Cognition* 20: 755–769.

Call, J. and Tomasello, M. (2011). 'Does the chimpanzee have a theory of mind? 30 years later', *Human Nature and Self Design* 83–96.

Cartmill, E. A. (2009). *Gestural Communication in Orangutans (Pongo pygmaeus and Pongo abelii): A Cognitive Approach* (Doctoral dissertation, University of St Andrews).

Cartmill, E. A. (2023). 'Overcoming bias in the comparison of human language and animal communication', *Proceedings of the National Academy of Sciences* 120(47): e2218799120.

Cartmill, E. A. and Byrne, R. W. (2010). 'Semantics of primate gestures: Intentional meanings of orangutan gestures', *Animal Cognition* 13: 793–804.

Cartmill, E. A. and Byrne, R. W. (2011). 'Addressing the problems of intentionality and granularity in non-human primate gesture', in G. Stam and M. Ishino (eds.), *Integrating Gestures*, 15–26. Amsterdam: John Benjamins.

Cartmill, E. A. and Hobaiter, C. (2019a). 'Developmental perspectives on primate gesture: 100 years in the making', *Animal Cognition* 22: 453–459.

Cartmill, E. A. and Hobaiter, C. (2019b). 'Gesturing towards the future: Cognition, big data, and the future of comparative gesture research', *Animal Cognition* 22: 597–604.

Cassell, J., McNeill, D., and McCullogh, K. E. (1999). 'Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information', *Pragmatic Cognition* 7: 1–33.

Clay, Z., Archbold, J., and Zuberbühler, K. (2015). 'Functional flexibility in wild bonobo vocal behaviour', *PeerJ* 3: e1124.

Crockford, C., Wittig, R. M., Mundry, R., and Zuberbühler, K. (2012). 'Wild chimpanzees inform ignorant group members of danger', *Current Biology* 22(2): 142–146.

Dafreville, M., Hobaiter, C., Guidetti, M., et al. (2021). 'Sensitivity to the communicative partner's attentional state: A developmental study on mother–infant dyads in wild chimpanzees (*Pan troglodytes schweinfurthii*)', *American Journal of Primatology* 83(12): e23339.

Darwin, C. (1872). *The Expression of Emotion in Man and Animals*. London: Murray.

ELAN (Version 6.4) [Computer software]. (2022). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive, https://archive.mpi.nl/tla/elan.

9780192886491_Book.indb 483 /12_first_proofs/first_pr…/xml_for_typesetting 27-Mar-25 23:04:54

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Fedurek, P., Slocombe, K. E., and Zuberbühler, K. (2015). 'Chimpanzees communicate to two different audiences during aggressive interactions', *Animal Behaviour* 110: 21–28.

Fröhlich, M., Kuchenbuch, P., Müller, G., et al. (2016a). 'Unpeeling the layers of language: Bonobos and chimpanzees engage in cooperative turn-taking sequences', *Scientific Reports* 6(1): 25887.

Fröhlich, M., Wittig, R. M., and Pika, S. (2016b). 'Should I stay or should I go? Initiation of joint travel in mother–infant dyads of two chimpanzee communities in the wild', *Animal Cognition* 19: 483–500.

Fröhlich M. and van Schaik, C. P. (2018). 'The function of primate multimodal communication', *Animal Cognition* 21: 619–629, https://doi.org/10.1007/s10071-018-1197-8.

Fuchs, M., Genty, E., Zuberbühler, K., and Cotofrei, P. (2023). 'ASBAR: An Animal Skeleton-Based Action Recognition framework. Recognizing great ape behaviors in the wild using pose estimation with domain adaptation', *bioRxiv* 09, https://doi.org/10.1101/2023.09.24.559236.

Genty, E. (2019). 'Vocal–gestural combinations in infant bonobos: New insights into signal functional specificity', *Animal Cognition* 22: 505–518, https://doi.org/10.1007/s10071-019-01267-0.

Genty, E. and Fuchs, M. (2023). GApS: A Coding Scheme for Great Apes Signals in ELAN (Version v1.0.2), https://greatapesgestures.github.io/.

Genty, E., Breuer, T., Hobaiter, C., and Byrne, R. W. (2009). 'Gestural communication of the gorilla (Gorilla gorilla): Repertoire, intentionality and possible origins', *Animal Cognition* 12: 527–546.

Genty, E., Clay, Z., Hobaiter, C., and Zuberbühler, K. (2014). 'Multi-modal use of a socially directed call in bonobos', *PloS one* 9(1): e84738.

Genty, E., Neumann, C., and Zuberbühler, K. (2015). 'Complex patterns of signalling to convey different social goals of sex in bonobos', *Pan Paniscus: Scientific Reports* 5: 16135, https://doi.org/10.1038/srep16135.

Goldin-Meadow, S., Wein, D., and Chang, C. (1992). 'Assessing knowledge through gesture: Using children's hands to read their minds', *Cognition and Instruction* 9: 201–219.

Goodwin, C. (2018). *Co-operative Action*. Cambridge: Cambridge University Press.

Graham, K. E., Hobaiter, C., Ounsley, J., et al. (2018). 'Bonobo and chimpanzee gestures overlap extensively in meaning', *PLoS Biology* 16(2): e2004825.

Graham, K. E., Furuichi, T., and Byrne, R. W. (2020). 'Context, not sequence order, affects the meaning of bonobo (*Pan paniscus*) gestures', *Gesture* 19(2–3): 335–364.

Graham, K. E., Badihi, G., Safryghin, A., et al. (2022). 'A socio-ecological perspective on the gestural communication of great ape species, individuals, and social units', *Ethology Ecology & Evolution* 34(3): 235–259.

Grice, P. (1989). *Studies in the Way of Words*. Boston, MA: Harvard University Press.

Grund, C., Badihi, G., Graham, K. E., et al. (2023). 'GesturalOrigins: A bottom-up framework for establishing systematic gesture data across ape species', *Behavior Research Methods* 1–16.

Heesen, R., Bangerter, A., Zuberbühler, K., et al. (2021). 'Assessing joint commitment as a process in great apes', *iScience* 24: 102872.

Heyes, C. M. (1998). 'Theory of mind in nonhuman primates', *Behavioral and Brain Sciences* 21(1): 101–114.

Higham, J. P. and Hebets, E. A. (2013). 'An introduction to multimodal communication', *Behavioral Ecology and Sociobiology* 67: 1381–1388.

9780192886491_Book.indb 484 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:04:54

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Hobaiter, C. and Byrne, R. W. (2011). 'The gestural repertoire of the wild chimpanzee', *Animal Cognition* 14: 745–767.

Hobaiter, C. and Byrne, R. W. (2014). 'The meanings of chimpanzee gestures', *Current Biology* 24(14): 1596–1600.

Hobaiter, C., Byrne, R. W., and Zuberbühler, K. (2017). 'Wild chimpanzees' use of single and combined vocal and gestural signals', *Behavioral Ecology and Sociobiology* 71: 1–13.

Hobaiter, C., Graham, K. E., and Byrne, R. W. (2022). 'Are ape gestures like words? Outstanding issues in detecting similarities and differences between human language and ape gesture', *Philosophical Transactions of the Royal Society B* 377(1860): 20210301.

Knox, A., Markx, J., How, E., et al. (2019). 'Gesture use in communication between mothers and offspring in wild orang-utans (*Pongo pygmaeus wurmbii)* from the Sabangau Peat-Swamp Forest, Borneo', *International Journal of Primatology* 40: 393–416.

Krupenye, C. and Call, J. (2019). 'Theory of mind in animals: Current and future directions', *Wiley Interdisciplinary Reviews: Cognitive Science* 10(6): e1503.

Leavens, D. A. and Hopkins, W. D. (2005). 'Multimodal concomitants of manual gesture by chimpanzees (*Pan troglodytes*): Influence of food size and distance', *Gesture* 5(1–2): 75–90.

Leavens, D. A., Hopkins, W. D., and Bard, K. A. (1996). 'Indexical and referential pointing in chimpanzees (*Pan troglodytes*)', *Journal of Comparative Psychology* 110(4): 346.

Leavens, D. A., Hopkins, W. D., and Bard, K. A. (2005). 'Understanding the point of chimpanzee pointing: Epigenesis and ecological validity', *Current Directions in Psychological Science* 14(4): 185–189.

Leavens, D. A., Russell, J. L., and Hopkins, W. D. (2010). 'Multimodal communication by captive chimpanzees (*Pan troglodytes*)', *Animal Cognition* 13: 33–40.

Levinson, S. C. and Holler, J. (2014). 'The origin of human multi-modal communication', *Philosophical Transactions of the Royal Society B: Biological Sciences* 369(1651): 20130302.

Liebal, K., Pika, S., and Tomasello, M. (2006). 'Gestural communication of orangutans (*Pongo pygmaeus*)', *Gesture* 6: 1–38.

Liebal, K., Waller, B. M., Slocombe, K. E., and Burrows, A. M. (2014). *Primate Communication: A Multimodal Approach*. Cambridge: Cambridge University Press.

Mielke, A., Badihi, G., Graham, K. E., et al. (2024). 'Many morphs: Parsing gesture signals from the noise', *Behavior Research Methods* 56(7): 6520–6537.

Partan, S. R. and Marler, P. (2005). 'Issues in the classification of multimodal communication signals', *The American Naturalist* 166(2): 231–245.

Pika, S., Liebal, K., and Tomasello, M. (2003). 'Gestural communication in young gorillas (*Gorilla gorilla*): Gestural repertoire, learning, and use', *American Journal of Primatology* 60: 95–111.

Pika, S., Liebal, K., and Tomasello, M. (2005). 'Gestural communication in subadult bonobos (*Pan paniscus*): Repertoire and use', *American Journal of Primatology* 65: 39–61.

Plooij, F. X. (1977). 'Some basic traits of language in wild chimpanzees?', in A. Lock (ed.), *Action, Gesture and Symbol*, 111–131. London: Academic Press.

Pollick, A. S., Jeneson, A., and de Waal, F. B. (2008). 'Gestures and multimodal signaling in bonobos', *The Bonobos: Behavior, Ecology, and Conservation* 75–94.

Premack, D. and Woodruff, G. (1978). 'Does the chimpanzee have a theory of mind?', *Behavioral and Brain Sciences* 1(4): 515–526.

Roberts, A. I., Roberts, S. G. B., and Vick, S. J. (2014). 'The repertoire and intentionality of gestural communication in wild chimpanzees', *Animal Cognition* 17: 317–336.

9780192886491_Book.indb 485 /12_first_proofs/first_pr/xml_for_typesetting 27-Mar-25 23:04:54

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Rodrigues, E. D., Santos, A. J., Hayashi, M., et al. (2021). 'Exploring greetings and leave-takings: Communication during arrivals and departures by chimpanzees of the Bossou community in Guinea', *Primates* 63: 443–461.

Rodrigues, E. D., Santos, A. J., Veppo, F., et al. (2022). 'Connecting primate gesture to the evolutionary roots of language: A systematic review', *American Journal of Primatology* 83(9): e23313.

Schel, A. M., Townsend, S. W., Machanda, Z., et al. (2013). 'Chimpanzee alarm call production meets key criteria for intentionality', *PLoS one* 8(10): e76674.

Scott-Phillips, T. C. (2016). 'Meaning in great ape communication: Summarising the debate', *Animal Cognition* 19(1): 233–238.

Slocombe, K. E. and Zuberbühler, K. (2007). 'Chimpanzees modify recruitment screams as a function of audience composition', *Proceedings of the National Academy of Sciences* 104(43): 17228–17233.

Slocombe, K. E., Waller, B. M., and Liebal, K. (2011). 'The language void: The need for multimodality in primate communication research', *Animal Behaviour* 81(5): 919–924.

Taglialatela, J. P., Russell, J. L., Pope, S. M., et al. (2015). 'Multimodal communication in chimpanzees', *American Journal of Primatology* 77(11): 1143–1148.

Tanner, J. E. and Byrne, R. (1996). 'Representation of action through iconic gesture in a captive lowland gorilla', *Current Anthropology* 37: 162–173.

Thompson, L. A. and Massaro, D. W. (1986). 'Evaluation and integration of speech and pointing gestures during referential understanding', *Journal of Experimental Child Psychology* 42(1): 144–168.

Tomasello, M., George, B. L., Kruger, A. C., et al. (1985). 'The development of gestural communication in young chimpanzees', *Journal of Human Evolution* 14(2): 175–186.

Tomasello, M., Call, J., Nagell, K., et al. (1994). 'The learning and use of gestural signals by young chimpanzees: A trans-generational study', *Primates* 35: 137–154.

Tomasello, M., Call, J., and Gluckman, A. (1997). 'Comprehension of novel communicative signs by apes and human children', *Child Development* 1067–1080.

Townsend, S. W., Koski, S. E., Byrne, R. W., et al. (2017). 'Exorcising Grice's ghost: An empirical approach to studying intentional communication in animals', *Biological Reviews* 92(3): 1427–1433.

Vail, A. L., Manica, A., and Bshary, R. (2013). 'Referential gestures in fish collaborative hunting', *Nature Communications* 4(1): 1765.

Waller, B. M., Whitehouse, J., and Micheletta, J. (2017). 'Rethinking primate facial expression: A predictive framework', *Neuroscience & Biobehavioral Reviews* 82: 13–21.

Wilke, C., Kavanagh, E., Donnellan, E., et al. (2017). 'Production of and responses to unimodal and multimodal signals in wild chimpanzees, *Pan troglodytes schweinfurthii*', *Animal Behaviour* 123: 305–316.

Wiltshire, C., Komedova, V., Cheetham, J. L., et al. (2023). 'DeepWild: Application of pose estimation tool DeepLabCut for behaviour tracking in wild chimpanzees and bonobos', *Journal of Animal Ecology*, https://doi.org/10.1111/1365-2656.13932.

Wittig, R. and Boesch, C. (2003). 'The choice of post-conflict interactions in wild chimpanzees (*Pan troglodytes*)', *Behaviour* 140(11–12): 1527–1559.

Zuberbühler, K. (2005). 'The phylogenetic roots of language: Evidence from primate communication and cognition', *Current Directions in Psychological Science* 14(3): 126–130.

9780192886491_Book.indb 486 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:04:54

CHAPTER 26

EVOLUTIONARY ROOTS OF HUMAN COOPERATIVE COMMUNICATION

Using the CA-assisted comparative approach for quantitative and qualitative analyses

SIMONE PIKA AND RAY WILKINSON

26.1 Introduction

Human sociality is characterized by an exceptional way of engaging in highly complex, mentally mediated, and extremely structured social interactions (Enfield and Levinson 2006). These are facilitated by humans' sophisticated cognitive skills (Tomasello 1999; Matsuzawa 2008), peculiar communication system—language (Hurford 2004)—and a level of cooperation suggested to be unique to humans only (Enfield and Levinson 2006; Rand and Nowak 2013). This capacity for social interactions also enables humans to participate in a common mental world where individuals have expectations about each other's behaviour and beliefs, about what we share and what we do not share concerning knowledge, intentions, and motivations.

It is such features of human sociality and social interaction which have in recent years been brought to the fore through an influential hypothesis—the 'interaction engine hypothesis' (Levinson 2006a, 2019). It also adds a new perspective to evolutionary scenarios of human language by postulating that it is not language that made human communication possible but rather a specific capacity for social interaction. This interactional ethology or 'cognition-for-interaction' (Levinson 2006b) antecedes language both phylogenetically and ontogenetically and involves specific social cognitive and interactional abilities. Social cognitive skills are, for instance, joint attention, common ground, collaboration, and reasoning about communicative intent (Clark 1996), and interactional skills concern face-to-face interaction, frequent employment of mutual gaze, and the exchange of rapid communicative turns. The latter, conversational

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

turn-taking (see, for other turn-taking systems, Sacks et al. 1974), is characterized by distinct organizational features, allowing (i) for one participant at a time to communicate or act (Sacks et al. 1974), (ii) to link turns into a sequence such that a turn can be seen as responsive to a prior turn rather than as unrelated to it, and (iii) to display understanding of a prior contribution by responding to it appropriately or, alternatively, displaying a difficulty in understanding it such that the original communicator may then repair, or redo, their message (Kitzinger 2012). There is some evidence that certain key organizational features of interaction, including aspects of turn-taking organization, can be thought of as candidates for universals across cultures and languages (e.g. Stivers et al. 2009; Dingemanse et al. 2015; Kendrick et al. 2020). However, further research into this area is warranted since currently investigated sample sizes across cultures are very small (Stivers et al. 2009). In addition, there is an ongoing debate, although with little relevant empirical data analysis, concerning cultural variation in turn-taking practices, especially in non-Western lifestyle groups (Weisz et al. 1984; Rogoff et al. 1993; Chaudhary 2012; Schroder et al. 2012).

Recently, the interaction engine hypothesis, and specifically precursors of involved skills and mechanisms and degrees of complexity, also started to inspire comparative researchers (e.g. Rossano 2013; Ben Mocha et al. 2019; Heesen and Frohlich 2022; Mondada and Meguerditchian 2022; Schel et al. 2022). In addition, the tremendous body of existing work on temporally coordinated signalling in non-human species (see, for recent reviews, Pika et al. 2018; Vanderhoff and Bernal Hoverud 2022; Ravignani et al. 2019) suggests that turn-taking and involved elements may indeed be one of the most ancient layers of the language system (Levinson and Holler 2014).

For instance, Pika and colleagues (2018) provided an overview of existing studies on temporally coordinated signalling with a special focus on types of signals used, time windows/temporal relationships involved, and the avoidance of overlap. Researching studies published over the last fifty years and across four animal taxa (insects, anurans, birds, and mammals), they showed that birds use a large variety of different signals, deploy flexible interaction strategies, and display relatively high levels of temporal precision during duetting (i.e. 'acoustic interactions between two partners of the opposite sex'; Pika et al. 2018 citing Bailey 2003) that seem to be tightly linked to a species' social structure. Concerning anurans, temporally coordinated signalling was predominantly used in the form of antiphonal advertisement calls (i.e. referred to as call-and-response with some authors applying the term to denote exchanges of the same call type) by males, characterized by overlap avoidance. However, in some species, overlap seems to be typical (e.g. American toad males) while other species also engaged in duets (e.g. American clawed frogs).

The overview of studies into temporally coordinated signalling in mammals showed that some signs of turn-taking seem to be present in all the major primate branches—prosimians, monkeys, smaller and great apes—and in four distantly related mammal groups—bats, cetaceans, elephants, and mole rats. The majority of studies concerned vocal exchanges and were characterized by considerable between-species variability with regards to involved temporal relationships. Recent studies have also been focusing

9780192886491_Book.indb 488 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:04:54

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

on gestural interactions and showed that some elements of human cooperative turn-taking are also present in non-vocal interactions.

Concerning insects, Pika and colleagues (2018) reported that temporal exchanges are very diverse and range from vibration, percussion, stridulation, click mechanisms, air expulsions, and bioluminescence.

Overall, Pika and colleagues (2018) concluded that despite all research efforts, the existing knowledge about temporally coordinated signalling of non-human animals in comparison to human social action during conversational turn-taking is currently relatively small, with methodological confounds preventing meaningful cross-species comparisons. The extent to which cooperative turn-taking is uniquely human or represents a homologous and/or analogous trait is, therefore, currently unknown (Pika et al. 2018).

To bridge these existing gaps and instigate systematic quantitative comparisons across species and taxa, Pika and colleagues adopted and expanded parameters used in the CA-assisted comparative approach (see for more details in section 26.1.1.) and formulated a new comparative framework. The framework focuses on four distinct elements that can be measured and reliably assessed:

(A) Flexibility of turn-taking organization
(B) Who is taking the next turn?
(C) When do response turns occur?
(D) What should the next turn do?

Element A—flexibility of turn-taking organization—refers to the flexibility of turn-taking production. This means that turn-taking sequences can be composed of different units, and can involve different modalities, sizes, and ordering of turns. Flexible production can also be shown by underlying cognitive flexibility, with individuals voluntarily changing and adjusting signals/actions to the attentional state of recipients, and the social context and matrices. Operationalization of this element may focus on assessing signal and action types involved (e.g. gestures, vocalizations, facial expression types and frequency, multimodal combinations), frequency, order, and sequence of involved units via means of capturing statistical association patterns (Bosshard et al. 2022; or hallmarks of intentionality Dennett 1971, Fröhlich et al. 2016).

Element B—who is taking the next turn?—centres on the interactants and their interactional behaviours, that is, who can or should produce the next signal and which techniques are used for allocating turns to individuals or parties. Operationalization of this element may focus on measuring gaze direction, distance between interactants, approach and signalling behaviour, and test for possible linkages between behaviours produced before the first turn and taking the turn.

Element C—when do response turns occur?—refers to the time window and temporal relationship between an initiating turn and its response turn. Operationalization of this element may focus on measuring the time windows between turns via 'units' (signal/gesture or action, turn one; Hilbrink and colleagues 2015, Kolff and Pika in

9780192886491_Book.indb 489 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:04:54

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

review), produced by one individual, and receiving a unit (signal/gesture or action) from another individual, involving role reversal (signaler/actor-recipient roles). If not reliable (e.g. the start of a gesture cannot be seen), it can be assessed by measuring three distinct parameters: delayed response (≥ 2 sec), immediate response (≥ 0 < 2 sec), and overlapping responses (< 0 sec) (Fröhlich et al. 2016).

Element D—what should the next turn do?—centres on one of the most fundamental structures in the organization of human conversational turn-taking: adjacency pairs (Schegloff 2007). Adjacency pairs have distinct characteristics, that is they can be recursively reproduced (Levinson 2013) and expanded in conversation. They consist—in a minimal, unexpanded form—of two turns, which are produced by different participants, are adjacently placed, and are relatively ordered into first pair parts (actions that initiate some exchange, e.g. requests), and second pair parts (responsive actions, e.g. grants) (Schegloff 2007). Operationalization of this element may focus on assessing order and sequence of turns involved, and frequency of adjacency pair-like sequences used via means of capturing statistical association patterns (Bosshard et al. 2022).

26.1.1 On which methods does this framework rest and how does it work?

The introduced comparative framework is based on the methodology of the CA-assisted comparative approach, which was developed in 2012 by Wilkinson and colleagues (2012). We will outline the methodology in terms of three aspects, which can be interlinked, i.e. data collection, data representation and transcription, and data analysis.

26.1.1.1 Data collection, data representation and transcription, and data analysis

As Mondada (2012: 32) noted: 'Conversation analysts use audio and video recordings of naturally occurring activities to study the details of action as they are temporally and sequentially arranged, moment-by-moment, by the participants within the very context of their activity'. A number of features here are relevant for how this methodology is applied within the comparative framework.

First, the aim is to capture 'naturally occurring' interactions, i.e. interactions which would have taken place, and taken place in the same way, if they were not being recorded for research purposes. The aim, therefore, is to minimize the potential impact of the recording device, and of any co-present researcher(s), on the interaction. Any possible interference (e.g. the animal displaying an awareness of the researcher's presence) is noted, and a decision has to be made as to whether that part of the recording constitutes ecologically valid data for the purposes of analysis.

Second, this focus on recording naturally occurring interactions and activities is linked to the aims of the analysis which include explicating how certain social practices (including vocalizations, signals, and other forms of embodied activity) are understood by the participants as constituting certain meanings or actions (such as requesting,

9780192886491_Book.indb 490 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:04:54

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

offering, greeting, etc.). These practices are assumed to be social in the sense that their meanings are understood by all competent members of the group, and this informs how the practices are produced and interpreted by the participants within natural social interactions. There is an assumption that social interaction is 'orderly at a minute level of detail' (Stivers and Sidnell 2012: 2) and that this is due to the fact that it is the product of contributions by participants who share common competences in reasoning about, and producing, contributions to interaction. An overarching goal of analysis is to describe the shared practices and their meanings/actions using, in the first instance, an inductive qualitative methodology (Stivers and Sidnell 2012).

Third, analysis focuses closely on how the participants act and react to each other within the interaction. Temporal and sequential issues are of central importance to how participants make sense to each other within a given interaction, and therefore are also important aspects of the analysis. For example, to understand what a particular vocalization or gesture 'means', analysis focuses on how the recipient of that vocalization or gesture appears to interpret it in the following turn. As Heritage and Atkinson (1984: 1) noted (in relation to the analysis of human social interactions), an advantage of this methodological focus on participants' behaviour is that '[t]he analyst is thus not required to speculate upon what the interactants hypothetically or imaginably understood.… Instead, analysis can emerge from observation of the conduct of the participants'. Thus, in analysing the meaning of a particular contribution, rather than attempting to hypothesize about a participant's intention, analysis can be based in part on the recipient's interpretation as seen through their response or reaction in the following turn.

The fact that close analytic attention is paid to the nature of the participants' vocal and embodied contributions, how these are reacted to in the following turns, and how these emerge within a particular sequential context of prior contributions, has a consequence for how data within this approach is transcribed and represented. Vocalizations and instances of embodied conduct are transcribed and/or represented (e.g. in the form of video stills or pictorial representations of these stills) not in isolation but rather in the context of what occurred within the interaction both prior to and following the focal contribution(s). In addition, since within CA no *a priori* assumptions are made by the analyst about which aspects of the participants' vocal or embodied conduct may be relevant (Heritage 1984), the transcription and representation of the data tends to be detailed (Wilkinson et al. 2012; Rossano 2013; Mondada and Meguerditchian 2022). In addition, close attention is paid to capturing the temporal and sequential unfolding of the interaction as it is produced by the participants' contributions on a moment-by-moment basis.

26.1.2 What is this method good for?

Over the last fifty years, CA has become the main approach within sociology and linguistics for investigating human social interaction (Stivers and Sidnell 2012). Based

9780192886491_Book.indb 491 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:04:54

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

on collections made up of detailed analysis of individual instances, CA has uncovered systematic organizational features of interaction. This approach has shown that participants draw on these organizational features to produce intelligible contributions, understand those contributions, and co-construct coherent sequences and larger episodes of linked contributions. These social organizations of practice constitute solutions to generic problems of how participants might coordinate their interactions (Schegloff 2006). They include, for human interaction, the organization of turn-taking, the organization of sequences of actions between participants whereby an initiating action by one participant, such as a request, is followed by a responding action by another participant, such as a granting or a declining of that request (Schegloff 2007) and the organization of repair (i.e. the practices whereby participants can highlight and potentially deal with errors in speaking or problems with hearing or understanding of another's speech; Kitzinger 2012).

The explication of these generic organizations of practice has a number of benefits for the issues addressed in this volume. First, it provides a detailed understanding of the nature of human social interaction, including aspects of turn-taking organization, sequence organization, and repair organization, which appear to be candidate universals across cultures (Stivers et al. 2009; Dingemanse et al. 2015; Kendrick et al. 2020). Second, it provides a methodology for the analysis of social interactions in non-human species (Wilkinson et al. 2012; Logue and Stivers 2012). Third, it provides a broad framework for systematically comparing interactions between taxa and species, enabling inferences about possible factors driving the onset and development of distinct communicative skills and underlying cognitive mechanisms. This provides for the possibility of also tackling in more detail the evolutionary trajectory of abilities composing the stratified system of language (Levinson and Holler 2014).

Regarding turn-taking specifically, CA has highlighted a number of features of this aspect of social interaction whereby participants can coordinate their talking together on a moment-by-moment basis and resolve 'who should talk or move or act next and when they should do so' (Schegloff 2006: 71). This highlighting of the importance of turn-taking in human social interactions and the fact that it is a precursor to language both ontogenetically and perhaps also phylogenetially (Levinson 2016) resulted in a considerable increase of turn-taking studies in other animal species.

While human co-participation in many interactional activities (e.g. dancing together) typically involves each participant moving or acting simultaneously with another, talking within social interaction is largely a one-at-a-time activity. As one participant talks, the other remains (relatively) silent (although embodied activity by a recipient, such as gesture or gaze at, or away from, the speaker, occurs, and can influence the emerging talk; Goodwin 2018). This one-at-a-time nature of talking within interaction has a consequence for the roles of the participants; one can, temporarily, be the primary active producer of a communicative act, while the other is primarily the recipient of that communicative act. Also, when (in a two-participant interaction) that recipient in the next turn-at-talk becomes the speaker, their contribution can be examined by the prior speaker/now recipient for how (and whether) this new turn-at-talk is *responsive* to the

9780192886491_Book.indb 492 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:04:54

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

prior turn (Schegloff 2006). Conversation is thus collaboratively co-constructed by the participants on a turn-by-turn basis.

In their influential paper on turn-taking in human conversation, Sacks and colleagues (1974: 700) noted some 'grossly apparent facts' about conversation, and attempted to produce a simple model which would be capable of accommodating them. As well as the one-speaker-at-a-time nature of conversation, these facts including observations both about the variability—or flexibility—of turns and also about the apparent orderliness of how turns are exchanged (e.g. that turn-taking commonly occurs with little overlapping talk). The resulting model of turn-taking consists of a component about the nature of the turns, a component about the allocation of the turns between speakers, and a series of rules linking the two components. One rule, for example, concerns how the next turn may be allocated by means of the current speaker producing a turn (e.g. a question) which selects another participant to be the next speaker (Element D of the comparative framework) (Schegloff 2007).

26.1.3 What have we learned from using this method?

As noted above, CA research on human social conduct is based on collections of instances from naturally occurring social interactions which together can provide evidence of relevant organizations of practice (e.g. turn-taking, repair, or sequences of actions) and how the participants produce their own conduct and interpret the conduct of others in relation to those organizations. A CA-assisted comparative approach can also make use of this method via microscopic (qualitative) but also macroscopic (quantitative) analyses. For instance, a recent study used the CA-assisted comparative approach to carry out quantitative analyses of interactions, by operationalizing distinct parameters (e.g. frequency of sequences used, measuring time windows, and adjacency pairs) thus enabling inferences and cross-species comparisons about key elements of turn-taking (Fröhlich et al. 2016).

In the following paragraphs, we will illustrate the microscopic, qualitative type of CA approach by examining part of an episode showing meat sharing in chimpanzees (Wilkinson et al. 2012). In our discussion of the example, we will highlight how, through using this methodology, the four key elements of turn-taking highlighted by the comparative framework can be illuminated (Pika et al. 2018).

26.1.3.1 Dataset and analysis

The chosen example stems from a dataset collected by SP in 2008, when investigating gestural communication in adult male chimpanzees of the Ngogo community, Kibale National Park, Uganda. The two chimpanzees in the extract we will focus on are Morten and Mwea. They are sitting in a tree and Morten is eating a recently hunted subadult red colobus monkey.

For reasons of space, we focus on four brief extracts only to illustrate the four key elements of turn-taking. The episode consists of three turn-taking episodes (Extracts

9780192886491_Book.indb 493 /12_first_proofs/first_pr…/xml_for_typesetting 27-Mar-25 23:04:54

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

1a, 1b, 1c) where a request from Mwea for some of the meat is followed by actions from Morten, who does not accede to the request. These requests take the form of two gestures, REACH OUT[1] towards the food and a PALM-UP gesture placed under the food (Wilkinson et al. 2012). Finally, in Extract 1d, Morten OFFERS a piece of the meat to Mwea, and Mwea takes it.

> **Extract 1a**. In the ~ 55 seconds preceding this episode, Mwea is displaying an interest in the meat by either gazing at it, moving to sit in front of Morten such that he is in his line of gaze while eating the meat, or REACHING OUT his hand towards the meat.

At the point we focus on, Mwea, who has moved to position himself in a particular participation framework (Goodwin and Goodwin 2004) in front of Morten while he eats, REACHES OUT once again towards the meat (Figure 26.1). This signal constitutes a distinct communicative turn (TURN 1) in the interaction, recognizable as a request for the meat.



FIGURE 26.1 Turn 1 of the sequence: Mwea (L) reaches out towards the meat.

To increase the understandability of the figures, the following visual mark-ups are used: a white line marks the position of Mwea's right arm, a grey line his left arm; a circle marks the position of the meat that Morten is holding; a dotted line marks where the object (arm or meat) is positioned, although it is not visible at that moment on the video; an arrow depicts movement, such as turning away, and its direction.

[1] From here, gestures are depicted in SMALL CAPITALS.

9780192886491_Book.indb 494 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:04:54

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Morten, however, proceeds to pull the meat away slightly to his left, turns away from Mwea, and continues to bite at it (Figure 26.2). In this sequential context, following a request where a granting or declining of the request is expectable, these actions are interpretable in a certain way, i.e. not just as a solitary behaviour, but as a particular response (TURN 2) in relation to the request, i.e. a declining to accede to it. Some evidence that this is Mwea's interpretation of Morten's action is evident in his reaction; he vocalizes, showing a bared-teeth display, leaning backwards, and presenting his genitals/belly (Figure 26.3), a display communicating the signaller's wishes of no harm or benign intent (Kim et al. 2022).



FIGURE 26.2 Turn 2 of the sequence: Morten (R) turns away from Mwea in response to his request.

Extract 1b. A few seconds after Extract 1a, Mwea again produces a communicative turn; while moving along the bough towards Morten, he produces a request for the meat by reaching towards it (TURN 1: Figure 26.4).

Again, however, Morten produces actions which are visible as actively declining to accede to that request (TURN 2). He turns to his left and at the same time moves the meat slightly to his left, away from the reaching hand of Mwea who is coming from the right (Figure 26.5). In addition, Morten's right hand moves up, slightly blocking Mwea's access to the meat.

9780192886491_Book.indb 495 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:04:55

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

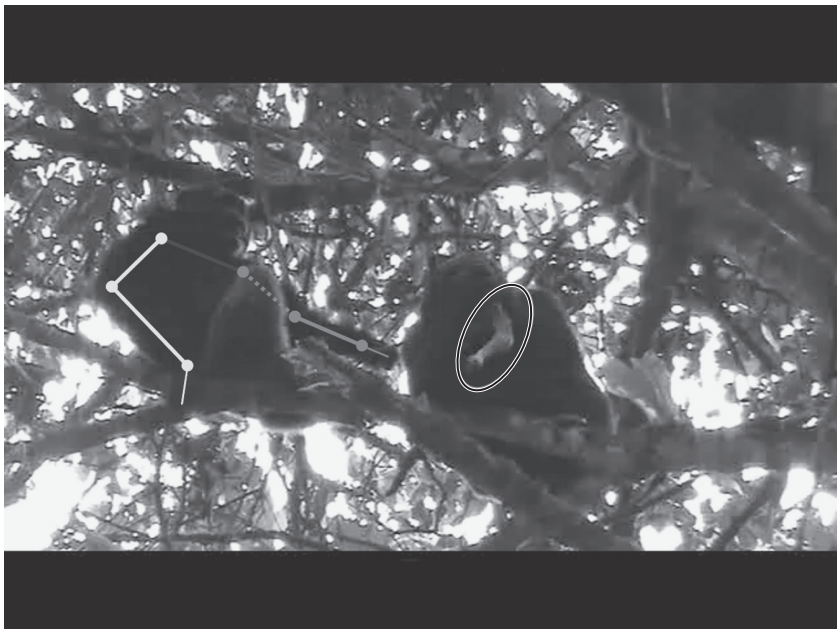


FIGURE 26.3 Mwea's (L) reaction to Morten's response: a bared-teeth display.



FIGURE 26.4 Turn 1 of the sequence: Mwea (L) REACHES towards the meat.

9780192886491_Book.indb 496 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:04:56

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

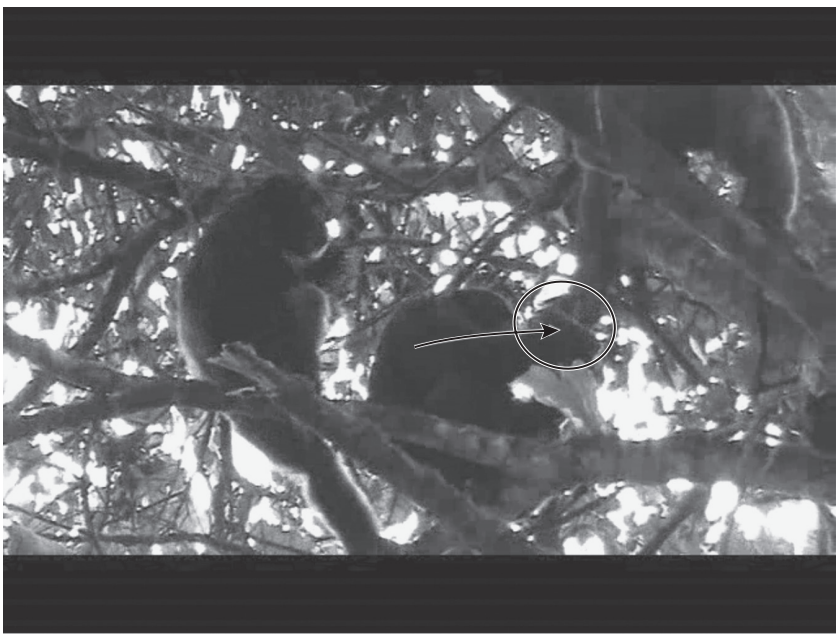


FIGURE 26.5 Turn 2 of the sequence: Morten (R) turns away and moves away the meat.

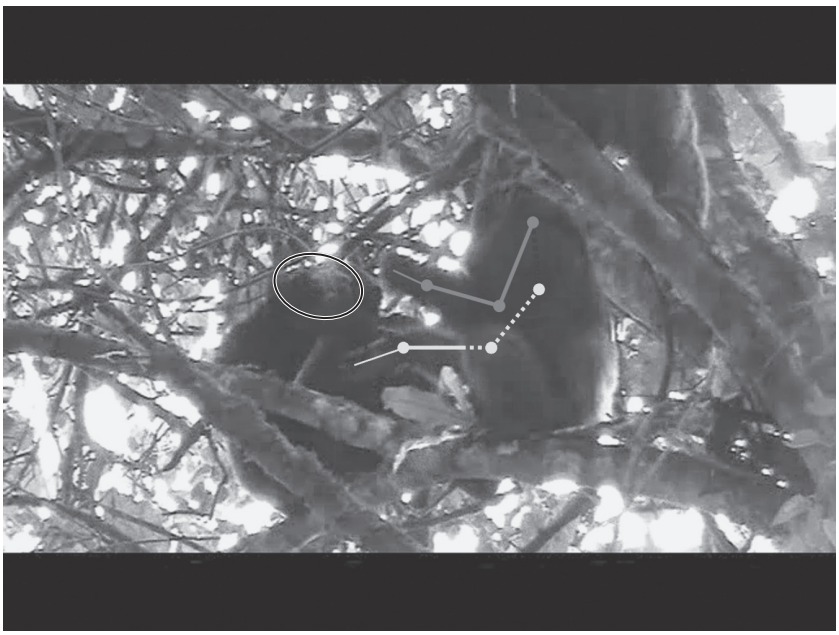


FIGURE 26.6 Turn 1 of the sequence: Mwea (R) requests meat with a PALM-UP gesture.

9780192886491_Book.indb 497 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:04:56

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Extract 1c. A few seconds after Extract 1b, Mwea moves past Morten on the branch and takes up a new participation framework in relation to Morten, facing him from Morten's left and positioning himself facing Morten with the meat between them. He then proceeds to again request some meat by reaching out and performing a PALM-UP gesture under it (TURN 1: Figure 26.6).

Morten again turns slightly away from Mwea to his right and moves the meat to the right away from Mwea. This behaviour is visible as Morten once more responding to Mwea's request by declining to accede to it (TURN 2: Figure 26.7).

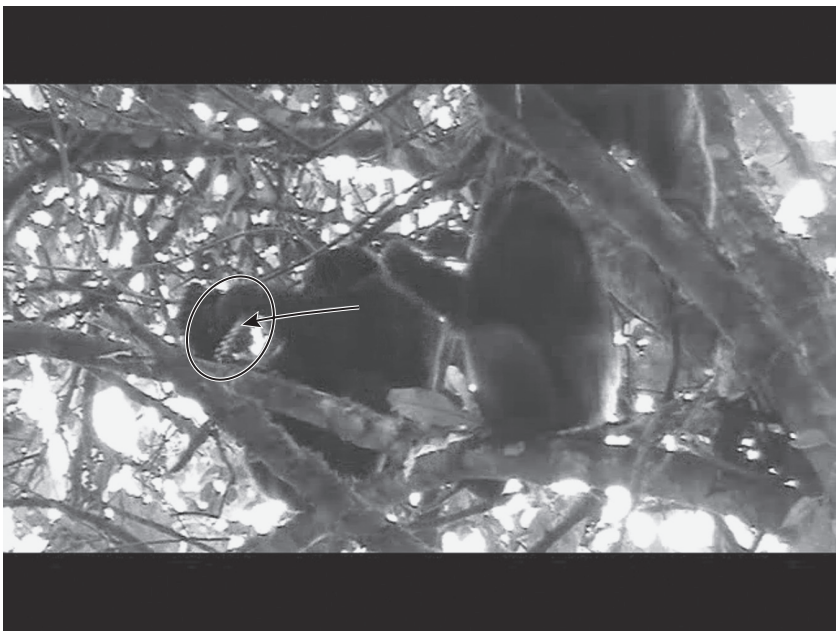


FIGURE 26.7 Turn 2 of the sequence: Morten (L) turns away and moves the meat away from Mwea.

Extract 1d. In the final part of this episode, which occurs ~ 25 seconds after Extract 1c, the meat is shared. Immediately preceding this, Morten starts to gnaw off a piece of the meat. Mwea appears to treat this as a possible precursor to the meat being shared; he moves along the bough past Morten and positions himself next to him, gazing down towards the meat. Morten then tears off a piece of meat and OFFERS it to Mwea using his left hand (TURN 1: Figure 26.8).

Mwea immediately responds to this OFFER by reaching down and taking the meat (TURN 2) from Morten. After the transfer of the meat, Mwea moves away up the branch carrying the meat.

9780192886491_Book.indb 498 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:04:57

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN



FIGURE 26.8 Turn 1 of the sequence: Morten (R) OFFERS the meat to Mwea.

26.2 Summary of the extract: Turn-taking behaviours in chimpanzee interaction

Extracts 1a–d show interactions characterized by the exchange of signals (e.g. reach out gesture) and actions (e.g. turn away), and display all four elements of the comparative framework (Pika et al. 2018). In terms of Element B (who is taking the next turn?) and D (what should the next turn do?), one chimpanzee produces an initiating signal (request gestures by Mwea in Extracts 1a, 1b, and 1c; an OFFER by Morten in Extract 1d). In each case, this constitutes the first part of an adjacency pair-type sequence, making expectable that the other chimpanzee will take the next turn and that this next turn will be a response to the initiating turn. Concerning Element C (when do the response turns occur), each case is characterized by a short time window between the initiating turn and the response. As regards Element A (the flexibility of the turn-taking organization), Mwea displays signal flexibility by using different signals to request meat (REACH OUT, PALM-UP) and, in parallel, to appease the meat owner (bared-teeth displays). He also displays flexibility concerning the position he takes up in relation to Morten (i.e. the different participation frameworks) such that he can, for example, place himself in Morten's line of vision and be close to the meat.

9780192886491_Book.indb 499 /12_first_proofs/first_pr…/xml_for_typesetting 27-Mar-25 23:04:57

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

26.3 Conclusion

In this chapter, we have discussed the CA-assisted comparative approach (Wilkinson et al. 2012) as a method for investigating the interactional conduct of animal species, in particular here turn-taking in wild-living chimpanzees, and for developing a framework that permits comparisons of interactional conduct across species. We have outlined the conversation analytic methodology of data collection, representation/transcription, and analysis which forms the basis for this framework, and have discussed what this method offers in terms of examining turn-taking behaviour. Finally, we illustrated this method by presenting an example of turn-taking in the form of meat sharing in chimpanzees. By using this method to examine turn-taking across species and involved complexity, we are able to shed detailed light onto an interactional niche that may have been a crucial mechanism during the cradle of language (Levinson 2006a, 2019).

References

Bailey, W. J. (2003). 'Insect duets: Underlying mechanisms and their evolution', *Physiological Entomology* 28: 157–174, doi:10.1046/j.1365-3032.2003.00337.x.

Ben Mocha, Y., Mundry, R., and Pika, S. (2019). 'Joint attention skills in wild Arabian babblers (*Turdoides squamiceps*): A consequence of cooperative breeding?', *Proceedings of the Royal Society B* 286: 20190147, doi:10.1098/rspb.2019.0147.

Bosshard, A. B., Leroux, M., Lester, N. A., et al. (2022). 'From collocations to callocations: Using linguistic methods to quantify animal call combinations', *Behavioral Ecology and Sociobiology* 76(9): 122, doi:10.1007/s00265-022-03224-3.

Chaudhary, N. (2012) 'Select 9. Negotiating with autonomy and relatedness: Dialogical processes in everyday lives of Indians', in H. M. J. Hermans and T. Gieser (eds.), *Handbook of Dialogical Self Theory*, 169–184. Cambridge: Cambridge University Press.

Clark, H. H. (1996). *Using Language*. Cambridge: Cambridge University Press.

Dennett, D. C. (1971). 'Intentional systems', *Journal of Philosophy* 68(4): 87–106.

Dingemanse, M. et al. (2015). 'Universal principles in the repair of communication problems', *PLoS ONE* 10: e0136100, doi:10.1371/journal.pone.0136100.

Enfield, N. J. and Levinson, S. C. (2006). 'Introduction: Human sociality as a new interdisciplinary field', in N. J. Enfield and S. C. Levinson (eds.), *Roots of Human Sociality, Culture, Cognition and Interaction*, 1–35. New York: Berg.

Fröhlich, M. et al. (2016). 'Unpeeling the layers of language: Bonobos and chimpanzees engage in cooperative turn-taking sequences', *Scientific Reports* 6: 25887, doi:10.1038/srep25887.

Goodwin, C. (2018). *Co-Operative Action*. Cambridge, MA: Cambridge University Press.

Goodwin, C. and Goodwin, M. H. (2004). 'Participation', in A. Duranti (ed.), *A Companion to Linguistic Anthropology* 222–244. Oxford: Wiley-Blackwell.

Heesen, R. and Fröhlich, M. (2022). 'Revisiting the human 'interaction engine': Comparative approaches to social action coordination', *Philosophical Transactions of the Royal Society B Biological Sciences* 377: 9, doi:10.1098/rstb.2021.0092.

9780192886491_Book.indb 500 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:04:57

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Heritage, J. (1984). *Garfinkel and Ethnomethodology*. Oxford: Polity Press.

Heritage, J. and Atkinson, J. M. (1984). 'Introduction', in J. M. Atkinson and J. Heritage (eds.), *Structures of Social Action: Studies in Conversation Analysis*, 1–15. Cambridge: Cambridge: University Press.

Hilbrink, E. E., Gattis, M., and Levinson, S. C. (2015). 'Early developmental changes in the timing of turn-taking: A longitudinal study of mother–infant interaction', *Frontiers in Psychology* 6(1492), doi:10.3389/fpsyg.2015.01492.

Hurford, J. R. (2004). 'Human uniqueness, learned symbols and recursive thought', *European Review* 12: 551–565, doi:10.1017/S106279870400047X.

Kendrick, K. H. et al. (2020). 'Sequence organization: A universal infrastructure for social action', *Journal of Pragmatics* 168: 119–138, doi:10.1016/j.pragma.2020.06.009.

Kim, Y., Vlaeyen, J. M. R., Heesen, R., et al. (2022). 'The association between the bared-teeth display and social dominance in captive chimpanzees (*Pan troglodytes*)', *Affective Science* 3: 749–760, doi:10.1007/s42761-022-00138-1.

Kitzinger, C. (2012). 'Repair', in T. Stivers and J. Sidnell (eds.), *The Handbook of Conversation Analysis*, 229–256. Oxford: John Wiley & Sons.

Kolff, K. and Pika, S. (in review), 'Turn-taking in grooming interactions of chimpanzees (*Pan troglodytes schweinfurthii*) in the wild: The role of demographic and social factors', *Animal Cognition*.

Levinson, S. C. (2006a). 'On the human "interaction engine"', in N. J. Enfield and S. C. Levinson (eds.), *Roots of Human Sociality: Culture, Cognition and Interaction Wenner-Gren Foundation for Anthropological Research, Symposium 134*, 39–69. New York: Berg.

Levinson, S. C. (2006b). 'Cognition at the heart of human interaction', *Discourse Studies* 8: 85–93, doi:10.1177/1461445606059557.

Levinson, S. C. (2013). 'Recursion in pragmatics', *Language* 89: 149–162.

Levinson, S. C. (2016). 'Turn-taking in human communication—Origins and implications for language processing', *Trends in Cognitive Sciences* 20: 6–14, doi:10.1016/j.tics.2015.10.010.

Levinson, S. C. (2019). 'Interactional foundations of language: The interaction engine hypothesis', in P. Hagoort (ed.), *Human Language: From Genes and Brain to Behavior*, 189–200. Cambridge, MA: MIT Press.

Levinson, S. C. and Holler, J. (2014). 'The origin of human multi-modal communication', *Philosophical Transactions of the Royal Society B Biological Sciences* 369: 9, doi:10.1098/rstb.2013.0302.

Logue, D. M. and Stivers, T. (2012). 'Squawk in interaction: A primer of conversation analysis for students of animal communication', *Behaviour* 149: 1283–1298, doi:10.1163/1568539X-00003031.

Matsuzawa, T. (2008). *Primate Origins of Human Cognition and Behavior*. Berlin, Heidelberg: Springer.

Mondada, L. (2012). 'The conversation analytic approach to data collection', in J. Sidnell and T. Stivers (eds.), *The Handbook of Conversation Analysis*, 32–56. Oxford: John Wiley & Sons.

Mondada, L. and Meguerditchian, A. (2022). 'Sequence organization and embodied mutual orientations: Openings of social interactions between baboons', *Philosophical Transactions of the Royal Society B Biological Sciences* 377: doi:10.1098/rstb.2021.0101.

Pika, S., Wilkinson, R., Kendrick, K. H., and Vernes, S. C. (2018). 'Taking turns: Bridging the gap between human and animal communication', *Proceedings of the Royal Society B Biological Sciences* 285: 9, doi:10.1098/rspb.2018.0598.

Rand, D. G. and Nowak, M. A. (2013). 'Human cooperation', *Trends in Cognitive Sciences* 17: 413–425, doi:10.1016/j.tics.2013.06.003.

9780192886491_Book.indb 501 /12_first_proofs/first_pr/xml_for_typesetting 27-Mar-25 23:04:57

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Ravignani, A., Verga, L., and Greenfield, M. D. (2019). 'Interactive rhythms across species: The evolutionary biology of animal chorusing and turn-taking', *Annals of the New York Academy of Sciences* 1453: 12–21, doi:10.1111/nyas.14230.

Rogoff, B. et al. (1993). 'Guided participation in cultural activity by toddlers and caregivers', *Monographs of the Society for Research in Child Development* i–179.

Rossano, F. (2013). 'Sequence organization and timing of bonobo mother-infant interactions', *Interaction Studies* 14: 160–189, doi:10.1075/is.14.2.02ros.

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). 'A simplest systematics for the organization of turn-taking in conversation', *Language* 50: 696–735.

Schegloff, E. A. (2006). 'Interaction: The infrastructure for social institutions, the natural ecological niche for language, and the arena in which culture is enacted', in N. J. Enfield and S. C. Levinson (eds.), *Roots of Human Sociality, Culture, Cognition and Interaction*, 70–96. London: Berg.

Schegloff, E. A. (2007). *Sequence Organization in Interaction: Volume 1: A Primer in Conversation Analysis*. New York: Cambridge University Press.

Schel, A. M., Bono, A., Aychet, J., et al. (2022). 'Intentional gestural communication amongst red-capped mangabeys (*Cercocebus torquatus*)', *Animal Cognition* 25: 1313–1330, doi:10.1007/s10071-022-01615-7.

Schröder, L., Kärtner, J., Keller, H., and Chaudhary, N. (2012). 'Sticking out and fitting in: Culture-specific predictors of 3-year-olds' autobiographical memories during joint reminiscing', *Infant Behavior and Development* 35: 627–634.

Stivers, T. and Sidnell, J. (2012). 'Introduction', in J. Sidnell and T. Stivers (eds.), *The Handbook of Conversation Analysis*, 1–8. Oxford: John Wiley & Sons.

Stivers, T. et al. (2009). 'Universals and cultural variation in turn-taking in conversation', *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* 106: 10587–10592, doi:10.1073/pnas.0903616106.

Tomasello, M. (1999). *The Cultural Origins of Human Cognition*. Cambridge, MA: Harvard University Press.

Vanderhoff, E. N. and Bernal Hoverud, N. (2022). 'Perspectives on antiphonal calling, duetting and counter-singing in non-primate mammals: An overview with notes on the coordinated vocalizations of bamboo rats (*Dactylomys* spp., Rodentia: Echimyidae)', *Frontiers in Ecology and Evolution* 10: 1–10, doi:10.3389/fevo.2022.906546.

Weisz, J. R., Rothbaum, F. M., and Blackburn, T. C. (1984). 'Standing out and standing in: The psychology of control in America and Japan', *American Psychologist* 39: 955.

Wilkinson, R., Leudar, I., and Pika, S. (2012). 'Requesting behaviors within episodes of active sharing: A new look on chimpanzee signaling', in S. Pika and K. Liebal (eds.), *Developments in Primate Gesture Research*, 199–221. Amsterdam: John Benjamins Publishing Company.

9780192886491_Book.indb 502 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:04:57

CHAPTER 27

MULTI-LEVEL APPROACHES TO ASSESSING FLEXIBILITY IN GREAT APE COMMUNICATION

MARLEN FRÖHLICH AND CAREL P. VAN SCHAIK

27.1 Adopting a multi-level comparative approach to decipher language origins

One defining feature of human communication is its remarkable plasticity, given that effective language use not only relies on highly flexible adjustments to socioecological settings, but also on the creation and comprehension of novel utterances with novel meanings (i.e. 'productivity'; Hockett 1960). The comparative approach of communication among extant non-human species has been considered a critical window to learn about selective pressures and cognitive demands leading to the emergence of human language (Hauser et al. 2002; Fitch 2010; Levinson and Holler 2014; Cartmill 2023). Most of the relevant comparative research has focused on the great apes, our closest living relatives: chimpanzees (*Pan troglodytes*), bonobos (*Pan paniscus*), gorillas (*Gorilla* spp.), and orangutans (*Pongo* spp.) (Call and Tomasello 2007; Liebal et al. 2013).

The communicative repertoires of non-human species are generally regarded as largely innate (Fitch 2020), and also rather small with respect to vocalizations (Arnold and Zuberbühler 2012; Seyfarth and Cheney 2014). In particular, for the vocal domain, it has been argued that plasticity is seriously lacking (Fischer et al. 2015; Fischer and Price 2017). However, for gestures, some flexibility in form and function has been documented (see section 27.1.3). Moreover, great apes tend to be remarkably innovative in other non-communicative domains (e.g. van Schaik et al. 2006; Manrique et al.

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

2013; Bandini and Harrison 2020). Both observations suggest that the extent to which primates can adjust their communicative behaviour in response to the immediate ('behavioural flexibility') and developmental environment ('ontogenetic plasticity') is less limited than often assumed, especially when it comes to great ape gestures. Their study should inform us about the degree of communicative innovativeness and creativity prior to the emergence of human language.

Progress in comparative research on the flexibility and plasticity in animal communication has been hampered by inconsistencies in terminology. These even concern basic concepts in the field, like 'communication', 'signal', and 'meaning' (Scott-Phillips 2008; Rendall et al. 2009; Ruxton and Schaefer 2011; Frohlich and van Schaik 2020). For example, evolutionary biologists consider as signals such acts or traits that have evolved because of their effect on others (Maynard Smith and Harper 2003; Scott-Phillips 2008). However, there is accumulating evidence for population- or even individual-specific meanings of social signals, that is, primate gestures and vocalizations (e.g. Fichtel and Van Schaik 2006; Halina et al. 2013; Badihi et al. 2023). This means that communicative acts that resemble signals with regard to both form and function (meaning) can also be acquired ontogenetically rather than evolutionarily inherited; they can be called learned (rather than evolved) signals (Fröhlich and van Schaik 2020). Their presence indicates at least some level of flexibility and thus communicative plasticity.

For both practical and ethical reasons, we are faced with substantial limitations regarding opportunities for controlled experiments in non-human primates, and particularly with the long-lived and highly social great ape species. Since it is rather difficult to elicit social signals or responses from them via non-invasive experiments (predator-induced alarm calls seem to be an exception, but generally yield only very small and often pseudo-replicated samples), researchers mostly rely on observational paradigms, recording instances of spontaneous intra-specific communicative interaction.

Long-term studies with the goal of estimating the extent of ontogenetic plasticity in great apes' communicative behaviour (e.g. studies with the goal of teaching sign language to enculturated great apes) are extremely challenging and also suffer from small sample sizes. Here, we argue that a more practical approach can be to study multiple levels of variation in naturalistic communication: populations, research settings, dyads, and individuals. In the following, we will briefly review the evidence for each of these levels of comparison.

27.1.1 Between-population comparisons

Analogous to the widely known attempts of identifying cultural variants in great ape behaviour (Whiten et al. 1999; van Schaik et al. 2003), one critical approach to assessing ontogenetic plasticity in great apes' communicative behaviour is to gather evidence on geographic variation in the use of vocalizations and gestures. Of course, this 'method of exclusion' can be applied not only to wild populations, but also to zoo-housed social groups. In 2012, Wich and colleagues described potential cultural call variants in several

9780192886491_Book.indb 504 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:04:58

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

populations of wild Bornean and Sumatran orangutans, which were apparently caused by the invention and socially induced imitation of novel sounds and vocalizations (Hardus et al. 2009; Wich et al. 2012). These included 'kiss-squeak' variants (a sound produced in distress, occasionally involving tool use), 'nest smacks' and 'raspberries' (both produced during nest-building), and 'throat scrapes' and 'harmonic uuhs' (both produced by mothers before retrieving their infant) (van Schaik et al. 2006). Likewise, in wild chimpanzees, sounds performed during grooming bouts (e.g. 'splutters' and 'teeth chomps'; Watts 2016) and local variants of leaf-modifying gestures (Badihi et al. 2023) seem to provide tentative evidence for sound and gestural dialects, respectively. A future large-scale project that systematically maps differences in repertoire and functional use across captive and wild study sites seems to be in order, and may offer further insights on the precursors of the productivity and immense plasticity of human language. If this form of behavioural plasticity is indeed limited to great apes (i.e. absent in less closely related species like monkeys), this has important implications for theories of language evolution.

27.1.2 Wild–captive comparisons

We recently proposed another population-level approach to estimate the extent of ontogenetic plasticity in great apes' communicative behaviour: systematically comparing groups or populations of the same species living in the wild and in humanmade, artificial habitats in captivity (Kummer and Goodall 1985; Frohlich and van Schaik 2020). Wild–captive contrasts may give crucial insights into how repertoires respond to radical differences in the social and physical environment. In captivity, individuals of most primate species are closer together and thus in more direct contact, are often more visible to each other, and more terrestrial than in the wild. Moreover, captive animals do not need to forage for their food and are exposed to behavioural enrichment, but virtually no predatory risk (Haslam 2013). Space in captive environments is obviously limited and less three-dimensionally complex, whereas substrate types are different from those in the wild, with less variation in heights, inclines, and increased hard surfaces (Lewton 2017).

Experiments with captive individuals are often governed by captivity bias, also known as the 'captivity effect' (Meulman et al. 2012; Haslam 2013). This bias is supported by studies reporting that captive animals, particularly primates, outperform their wild counterparts in certain tasks such as tool innovation and problem-solving. Studies on curiosity and problem-solving in orangutans revealed a striking wild–captive contrast, with captive individuals generally being much less neophobic than their wild counterparts (Forss et al. 2015; van Schaik et al. 2016). Thus, the captivity-specific conditions introduced above, including enhanced terrestriality, food availability, and lack of predators, might eventually lead to higher innovation rates, especially when food is provisioned (Leca et al. 2010). This is indeed commonly observed (Kummer and Goodall 1985; Reader and Laland 2003; Lehner et al. 2010; van Schaik 2016), and

9780192886491_Book.indb 505 /12_first_proofs/first_pr…/xml_for_typesetting 27-Mar-25 23:04:58

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

generally thought to reflect the combination of lower constraints (no predation, no foraging) and greater opportunities (loss of neophobia, behavioural enrichment).

The same effect holds in the communicative domain, where we now see direct evidence for such plasticity (for review, see Fröhlich and van Schaik 2020). In addition to increased opportunities, signal inventions may additionally reflect communicative necessities because of the restricted space and lack of choice of associates in human-made settings. Indeed, captivity generally elicits more repeated social interactions with the same individuals, but also with different sets of interactants than in the wild. For example, in some captive settings, chimpanzees use several invented, attention-getting sounds or vocalizations aimed at human caretakers (Hopkins et al. 2007), which are apparently transmitted from mother to offspring (Taglialatela et al. 2012). Wild–captive comparisons of primates' communicative behaviour are still rare, thus we recently carried out the first systematic study in orangutans (see case study presented in section 27.3.1; Fröhlich et al. 2021a, b).

27.1.3 Between-dyad comparisons

There is lively debate about what causes behavioural flexibility and developmental plasticity in non-human primates, and why it appears to be greater in great apes. One could argue that the fission–fusion social structures that characterize humans and other ape species, and the concomitant variation in the identity of interaction partners makes the context for social interaction and the response more unpredictable than in stable groups with a more rigid social structure (Freeberg et al. 2019). These circumstances require individuals to flexibly adjust their communications according to rapidly changing social challenges, a phenomenon which has been called 'social competence' (Taborsky and Oliveira 2012) or 'reversible phenotypic plasticity' (Dingemanse and Wolf 2013). Great apes are well known to coordinate social interactions via highly flexible communicative strategies, with signal use thus varying according to age difference, sex combination, and familiarity between signaller and recipient (Genty et al. 2015; Frohlich et al. 2020b; Graham et al. 2020). Our research on wild chimpanzees and orangutans has demonstrated that great apes attune their communicative signals to age, sex and status of, and the kin relationship with, their interaction partner (e.g. Fröhlich et al. 2016a, 2020b, 2021b). With increasing familiarity and social tolerance linked to predictable outcomes, as between mother and offspring, communicative exchanges appear to become much subtler (Fröhlich et al. 2016a). Elaborate communication involving persistence and multiple sensory components, on the other hand, may be more likely to be observed beyond the mother–infant pair. For instance, a chimpanzee infant may predominantly deploy mechanically effective actions to directly achieve their goals (e.g. taking the food held by mother) rather than produce communicative signals to indirectly achieve their goals (e.g. begging for food held by others; Fröhlich et al. 2020). This means that displays of communicative plasticity may be more readily observed in some interaction dyads compared to others.

9780192886491_Book.indb 506 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:04:58

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Moreover, if different dyads use different signal types for the same purpose, even within the same social group, this has important implications for mechanisms of acquisition (i.e. genetic predisposition vs. individual learning). Specifically, dyad-specific signals may point towards individual learning mechanisms, such as Ontogenetic Ritualization (OR) or 'social negotiation' (Pika and Fröhlich 2019). The OR theory proposes that gesture types are derived directly from repeated social interactions in which individuals participate through an individual learning process, and are thus acquired during an individual's lifetime (Tomasello 1990; Tomasello et al. 1994). Consequently, OR used to be the most compelling argument of why gestures are learnt rather than innate and thus fundamentally distinct from other communicative acts (Tomasello 2008). Taken together, it is important not only to take into account *that*, but also to explicitly study *how*, communicative behaviour differs between various interactional pairings that form in great ape social groups.

27.1.4 Between- and within-individual comparisons

Although the behaviour in apes' day-to-day communicative exchanges is highly flexible, certain signalling features may vary consistently between individuals (Fröhlich and Hobaiter 2018), and individuals may also differ in the extent of their communicative flexibility (Dingemanse et al. 2010). Previous work on behavioural coping styles in animals (i.e. correlated sets of individual behavioural traits that are consistent over time and contexts) has shown that individual differences in average behavioural expression and behavioural flexibility are often interlinked, and that individual differentiation in behavioural flexibility is adaptive (Koolhaas et al. 1999). If the number of individuals and samples is sufficiently large, repeated observations of communicative interactions in the same individuals in different dyadic constellations and social contexts can be analysed using a 'behavioural reaction norm approach' (Dingemanse et al. 2010): partitioning behavioural variability into intrinsic among-individual variation and reversible behavioural plasticity (i.e. environmental components).

Among all the levels of variation mentioned here, the individual level has probably been neglected most in comparative research on great apes. Unexplained variation around the mean is essentially viewed as 'noise', which is assumed to disappear once we gather sufficient data. While recent years have seen more efforts to tease apart between- from within-individual variation in great apes' social interactions (Tkaczynski et al. 2020; Frohlich et al. 2022), the focus of comparative research in general is still clearly on species- and population-level variation. Importantly, a focus on individuals can help to assess the role of plastic responses to the immediate (i.e. behavioural flexibility) and developmental environment (ontogenetic plasticity) in great ape communication and understand whether such flexibility has the potential to evolve.

9780192886491_Book.indb 507 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:04:58

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

27.1.5 Limitations of comparative observational paradigms

As early as 1967, Peter Marler, one of the pioneering comparative communication researchers, noted:

> There is great difference between setting out to understand a new language and embarking on an analysis of nonhuman communication. With animals, there is no choice but to approach the analysis as a nonparticipant. In this circumstance, clues to the occurrence of communication between two animals must be found in changes in the behaviour of one upon its perception of a signal from the other. (Marler 1967: 769)

It is therefore self-evident that purely observational studies come with several caveats, especially if data collection takes place in highly different research settings (Fröhlich et al. 2021a). First, wild and captive environments diverge profoundly in conditions for behavioural observation, with sampling of sufficient high-quality data on subtle visual or acoustic signals hampered by dense vegetation in the wild, and by mesh (visual obstruction) or glass barriers (acoustic obstruction) in the zoo. Second, comparative researchers working in both research settings typically need high-quality video and audio footage since it is virtually impossible to document communicative behaviour in real time (e.g. Genty et al. 2009; Hobaiter and Byrne 2011; Heesen et al. 2020; Frohlich et al. 2021a). Depending on recording intensity, this footage can be accompanied only by instantaneous sampling protocols that allow for complementary data on general activity, association patterns, and environmental variables (Fröhlich et al. 2017). Compared to experimental, more standardized paradigms, extensive sampling efforts are necessary in order to avoid bias with regard to individual and demographic variables. Previous work has shown that as much as sixty hours of observation per individual may be necessary to approach or reach asymptotes of cumulative signal repertoires (Hobaiter and Byrne 2011; Halina et al. 2013; Frohlich et al. 2016b). Third, while the data collection procedure can already be considered time-consuming, this applies even more to data annotation (i.e. video coding), which mostly takes many times over the duration of rewatching the actual video-recorded interaction. Lengthy coding procedures may well be the biggest drawback of the paradigm proposed here (although this applies to all observational studies focusing on subtle behavioural patterns). Establishing acceptable intercoder reliability (e.g. Cohen's Kappa values above 0.7; see Bakeman and Quera 2011) may take a long time, and coding all the relevant source material will take considerably longer. Dividing the coding work into smaller chunks for multiple coders in order to accelerate coding progress not only incurs further financial costs for research projects, but also increases risks of introducing observer-specific biases. Inevitably, we have to face a trade-off between coding effort, coding accuracy, and information content of the final dataset (Grund et al. 2023). The long anticipated development of automated coding techniques will hopefully soon finally lift this constraint.

Nonetheless, once we end up with a fully coded dataset, we have countless opportunities for analysing communicative behaviour. In most cases, behavioural observations of intraspecific interactions in naturalistic social contexts yield many

9780192886491_Book.indb 508 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:04:58

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

more data points than controlled laboratory experiments. With the advance of statistical techniques that control for complex confounding ecological and social variables, human control over the context as enabled by laboratory context may become less critical than it used to be. Mixed-effect modelling approaches (e.g. Bolker et al. 2009; Dingemanse and Dochtermann 2013; Bates et al. 2014) now allow us to tease apart effects of key test predictors like research setting or interaction partner from those of relevant control variables, like individual or group identity, age, sex, and so on. Thus, with an observational comparative approach, we benefit from data on naturalistic, spontaneous communicative interactions among conspecifics while, at the same time, controlling for biologically meaningful variation in individuals and the environment.

27.2 What can we learn from a multi-level approach applied to great ape communication?

To unravel whether the evolutionary roots of the remarkable plasticity in our communication system can be found in the great ape (hominid) lineage, we can ask which specific social and ecological conditions bring about a tendency in individuals to modify or even invent communicative acts and structures, and in turn also comprehend such novel acts. In other words: what triggers communicative innovativeness in non-human species, and which factors constrain this form of phenotypic plasticity? In our view, this question is highly relevant to any debate on language origins, and to make progress on answering it, we must distinguish innate from socially learned communicative acts while considering species-specific differences in social tolerance. To quantify the adjustment to immediate context as well as to the socioecological environment, we need data on variability revealed through comparisons of signal production and use between species which differ in social tolerance, but also between populations, settings, dyads, and individuals (Fröhlich et al. 2022; Graham et al. 2022). Behavioural variables of interest include, but are not limited to: repertoire size, repertoire composition, multimodal use of signals, function (e.g. 'apparently satisfactory outcomes'; Hobaiter and Byrne 2014), functional specificity, response latency, and communicative repair reflected in persistence and elaboration. The extent of variability in any of these patterns should reflect the degree to which individuals adjust to the specific socioecological environments in which they interact and grow up in.

The predictions for the extent and direction of variation in the use of communicative signals may differ strongly between study species. For example, in semi-solitary and arboreal species like orangutans, social interactions are much more frequent and take place more on the ground in captive settings (e.g. Zucker et al. 1978; Maple 1980; Kopp and Liebal 2018; Frohlich et al. 2021b), which may trigger the production of additional signals that are completely absent in the wild. In other words, conditions in captivity may

9780192886491_Book.indb 509 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:04:58

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

enable the production of signals not feasible or useful in natural environments (Marler 1965), and we would particularly expect a surge of signals whose production relies on flat substrates, and in behavioural contexts that are not common in wild settings (e.g. play, grooming, aggression, and mating). These larger communicative repertoires then result from the emergence of 'weak innovations' (Ramsey et al. 2007), which are widespread in the great ape behaviour repertoire as they can be readily reproduced by multiple individuals when conditions are suitable (Lehner et al. 2010; Tennie et al. 2020).

These predictions are not necessarily the same for other great apes. In chimpanzees, the wild–captive contrast may be reversed because their sociality in natural environments is characterized by complex fission–fusion dynamics. Wild chimpanzees have large territories that vary both between and within subspecies (e.g. 6.4–36.6 km^2 in western chimpanzees; Lemoine et al. 2020), and the types of arboreal vs. terrestrial substrates (Doran and Hunt 1996). Chimpanzees engage in large territory patrols (Mitani and Watts 2005) and large-scale hunting (Boesch and Boesch 1989). One would thus expect chimpanzee communications in captivity to be much less varied and more restricted than those of their wild counterparts. For gorillas, fewer differences between captive and wild groups may be expected, since they live in a stable, cohesive social system with group sizes that may not differ too much between the two research settings (Robbins et al. 2004). On the species level, differences in gesture form can be expected between the arboreal wild orangutans and the semi-terrestrial chimpanzees regarding the limbs involved in gesture production (Knox et al. 2019). Thus, a species' natural history will inform comparative researchers' predictions for the directions of the wild–captive contrast and thus the extent of plasticity to be expected for the respective communication system.

27.3 Insights gained using a 2 x 2 comparative approach for orangutan communication

27.3.1 Wild–captive contrasts in orangutans' non-vocal communication

We recently adopted a 2 x 2 approach (captive vs. wild; Borneo vs. Sumatra) in the first systematic study on wild–captive contrasts in orangutans' non-vocal communicative repertoires, that is, gestures and facial expressions (Fröhlich et al. 2021a). In our view, this great ape genus is exceptionally suited for this approach, because the captive–wild difference in socioecology is arguably greater than for any other hominid taxon. In the wild, orangutans live an arboreal and semi-solitary life, which is distinct from the social and terrestrial conditions they are exposed to in zoo settings, where visual

9780192886491_Book.indb 510 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:04:58

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

communication is less hampered by arboreality and obscuring vegetation. Moreover, association frequency is higher in the well-studied (Northwest-) Sumatra (i.e. Suaq and Ketambe) compared to Bornean orangutan populations, presumably enabled by higher fruit availability (van Schaik 1999; Fox et al. 2004; Mitra Setia et al. 2009; Wich et al. 2011). Thus, Bornean orangutans are clearly situated at the solitary end of the fission–fusion spectrum of mammals, while northwestern Sumatran populations (i.e. Suaq) reach sociability levels similar to those reported for certain communities of eastern chimpanzees (Boesch et al. 2002). Further studies showed that this population difference is also reflected in the number of socially learnt behavioural ('cultural') variants (van Schaik et al. 2003; Wich et al. 2012) and immatures' social interaction rates, including play and food begging (Fröhlich et al. 2020a). Thus, we can compare populations within the same genus that show a clear difference in social tolerance and gregariousness. In addition, orangutans' ability in elaborate and flexible signal use exhibited in captivity is certainly not inferior compared to that of other great apes (Liebal et al. 2006; Cartmill and Byrne 2007, 2010; Wich et al. 2009; Waller et al. 2015; Lameira et al. 2016).

Intriguingly, studying two species of orangutans that differ in social tolerance in both research settings permitted teasing apart the effect of intrinsic differences in social tolerance from wild–captive contrasts on signal production. Wild Sumatran orangutans were studied in Suaq, a field site in northern Sumatra, and wild Bornean orangutans in Tuanan, central Kalimantan, Borneo (for details, see Fröhlich et al. 2021a). Captive orangutans were studied in five different European zoos (Apenheul, Cologne, Munich, Munster, Zurich).

27.3.2 Repertoire size, repertoire composition, and functional specificity

After establishing signal repertoires and functions (i.e. presumed goals of communicative acts, with outcomes that apparently satisfied the signaller; see also Cartmill and Byrne 2010, Hobaiter and Byrne 2014), we tested several predictions about how the research environment affects repertoire sizes and functional specificity. Rather than coming up with a repertoire from scratch, we classified gestures and their outcomes by integrating previous work on zoo-housed and wild orang-utans (MacKinnon 1974; Rijksen 1978; Liebal et al. 2006; Cartmill and Byrne 2010; Knox et al. 2019). As an important initial step, we then checked whether cumulative plots of gesture types reached asymptotes, indicating that complete repertoires in the study groups were detected, at least for the contexts in which we were able to observe communicative interactions. This was the case for both wild (for which it was more relevant, since we predicted lower values for this setting) and most captive populations (see Figure 27.1).

9780192886491_Book.indb 511 /12_first_proofs/first_pr…/xml_for_typesetting 27-Mar-25 23:04:58

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

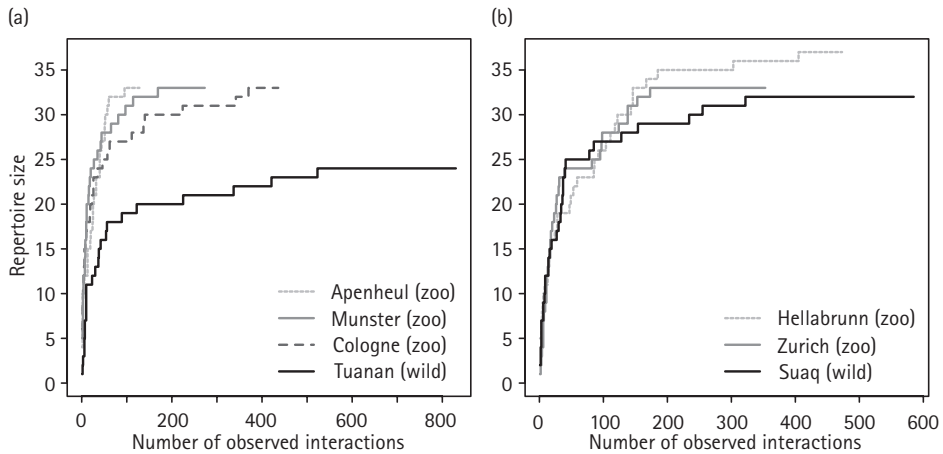


FIGURE 27.1 Cumulative number of identified signals in (a) Bornean and (b) Sumatran orangutans over consecutively observed interactions.

Asymptotes are depicted separately for the captive groups (a: Apenheul, Cologne, Munster, b: Hellabrunn, Zurich) and wild populations (a: Tuanan, b: Suaq) in this study.

Source: Fröhlich et al. (2021a).

We then identified a total of forty-one distinct non-vocal signal types, i.e. gestures and facial expressions. For the overall repertoire, we found nine signal types that were exclusively found in captive settings, vs. two only in the wild (Figure 27.2). This means around a quarter of the total orangutan non-vocal signalling repertoire was only observed in captivity.

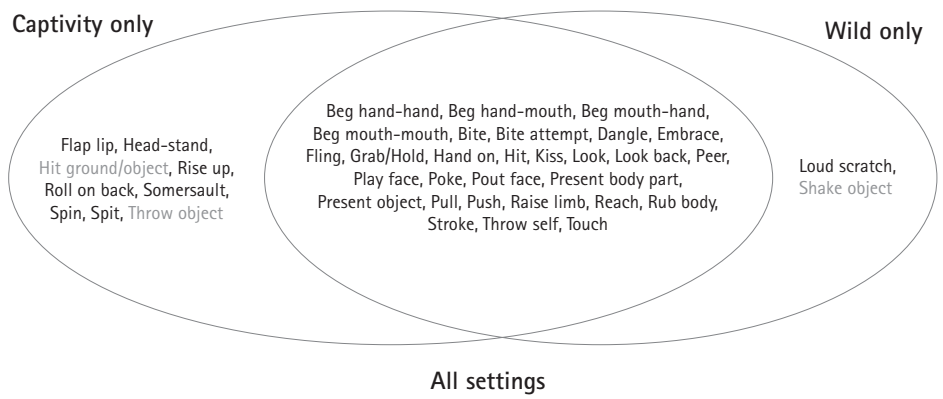


FIGURE 27.2 Overview of signals observed in wild and/or captive settings.

Signals documented by other studies in the contrasting setting are marked in grey.

Source: Fröhlich et al. (2021a).

Next, we looked at wild–captive contrasts in repertoire sizes on the individual level. Specifically, we fitted mixed models to test effects of research setting and orangutan species on individual repertoire size while controlling for age, sex, and sampling effort.

9780192886491_Book.indb 512 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:04:59

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Our first finding was that captive individuals have larger communicative repertoires irrespective of species, reflecting the increased opportunities for signal use in humanmade surroundings.

Second, we investigated whether the composition of individual repertoires systematically differed between captivity and the wild. To this end, we calculated similarity coefficients and conducted permutation tests revealing the extent of repertoire overlaps. Specifically, we were interested in whether the average repertoire overlap between two individuals living in the same research setting (e.g. captivity) differs from similarity between individuals living contrasting settings, separately for mothers and young immatures (because they provided the largest samples and are constant between settings), and separately for Bornean and Sumatran orangutans. We found that similarity of communicative repertoires was significantly larger within settings than between them, for both age classes in both species, suggesting that individuals respond to the very different affordances of captive vs. wild habitats.

Finally, we studied the functional specificity of signal types, that is how tightly communicative acts were linked to specific interaction outcomes (e.g. a social play initiation). Here, we also found a significant difference between captive and wild settings, but only for Sumatrans, the more sociable species. Apparently, the additional signals observed in zoo-living orangutans are also functionally more specific. A follow-up study confirmed that wild–captive contrasts in communicative repertoires seem to be mainly driven by further interaction partners and not by an increase in the number of social goals or interaction outcomes (Fröhlich et al. 2024). Specifically, immatures' repertoires expressed in the wild and in captivity varied significantly for signals directed at other individuals, not mother-directed signals or the number of social goals that immatures communicated towards.

Taken together, this study revealed pronounced wild–captive contrasts in non-vocal repertoire size on the individual and population level, as well as in the functional specificity of signal types. Orangutans facing more sociable and terrestrial conditions thus exhibit a noteworthy plasticity in their communicative behaviour, in that they produce further innate or innovated signals not recorded in naturalistic research settings. Since observation efforts in the wild were seemingly sufficient, the production of additional innate signals in captivity may result from habitat or context specificity that is absent from the sampled field sites. This can only be partly excluded (e.g. social play is common in the wild, too, but not the existence of flat substrate), thus the extent of genetic predisposition cannot be unambiguously determined. It is also worth noting that many of these 'captive only' and 'wild only' gestures are communicative acts that were previously observed in at least one group of another great ape species (e.g. shake object, somersault) and are thus more likely to have an innate component, but others seem to be good candidates for signal invention given that they were produced by single individuals only (e.g. spit, flap lip). However, the diversity of signal structure in animals arises not only to elicit different responses in recipients, but also reflects selective use towards the various classes of respondents that are available. With the emergence of a complex social life in which different group members have different roles to play, as we see may see in certain

9780192886491_Book.indb 513 /12_first_proofs/first_pr/xml_for_typesetting 27-Mar-25 23:04:59

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

zoo-housed groups of orangutans, the specification of appropriate respondents may require a considerable increase in the number of signal types used by a species (Marler 1967). Importantly, this effect is independent of any increase in the number of response patterns ('interaction outcomes') that may occur (Fröhlich et al. 2024).

27.3.3 Wild–captive contrasts beyond signal repertoires

The potential gain of insights offered by wild–captive comparisons is not restricted to communicative repertoires. Additional language-relevant aspects that were addressed with this large orangutan dataset included multimodality (Fröhlich et al. 2021b), communicative repair (Fröhlich and van Schaik 2022), as well as the extent of between- vs. within-individual variation (Fröhlich et al. 2022). For example, in a study focusing on social action coordination, we looked at wild–captive and Sumatran–Bornean contrasts in gestural redoings, that is the repetition and elaboration (i.e. modification) of initial gestures after communicative failure (Fröhlich and van Schaik 2022). We identified wild–captive contrasts in gestural elaboration in Borneans, but not in Sumatrans, the more socially tolerant species. Moreover, we found that elaboration was more effective in eliciting responses in Sumatrans, especially the captive ones, whereas the effectiveness of mere repetition was not influenced by species or setting. Since social tolerance is considered essential for extended social interactions, we argued that this feature drives the emergence of complex communicative exchanges in great apes.

27.4 Implications for studies of language evolution

Although the evidence for communicative plasticity presented here stems from orangutans, and the geographic variation in nesting and calling-infant sounds (van Schaik et al. 2006; Wich et al. 2012) has not been described for any of the African great apes, we have strong reason to believe that the extent of behavioural flexibility or plasticity in the communicative domain applies to whole hominid family. All great ape species have extensive signal repertoires to navigate their social lives, with gestural repertoire sizes exceeding those of monkeys (Call and Tomasello 2007; Byrne et al. 2017; Crockford 2019; Schel et al. 2022; but see Molesti et al. 2020), and their degree of behavioural innovation (Manrique et al. 2013; Bandini and Harrison 2020; Motes-Rodrigo and Tennie 2022) is unparalleled in the primate order.

Studies investigating contrasts on the population, group, dyadic, and individual levels can provide evidence for two different forms of phenotypic plasticity in communicative behaviour: behavioural flexibility as an *immediate* response and ontogenetic plasticity as a *long-term* response to the environment. On the immediate level, the

9780192886491_Book.indb 514 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:04:59

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

individual can communicate flexibly by using one and the same communicative act for several different functions, relying on the social context for disambiguation. On the long-term level, the intentional agent may begin to produce a novel communicative behaviour—either from its species-specific repertoire, previously unexpressed due to environmental constraints, or, if not available, produce one for which naive recipients (over repeated instances of interaction) infer its meaning from context and reactions.

Addressing the question of whether primate communicative behaviour is sufficiently plastic to generate signals from scratch, thus complementing species-typical repertoires, has important implications for scenarios of language evolution (see also Fröhlich and van Schaik 2020). Our ability to create and understand novel utterances with novel meanings (i.e. 'productivity'; Hockett 1960) reflects extreme communicative plasticity, the roots of which may be found in the hominid lineage. Our findings of behavioural plasticity in orangutan communication, but also the accumulating data on geographic variation in great apes' signal production and between-individual variation in plasticity, are consistent with the view that great apes, arguably more so than other primates, possess a latent capacity for communicative innovativeness or 'creativity'. If communicative plasticity did indeed evolve early on in the great ape lineage, our hominin ancestors could readily seize the opportunity to use invented expressions once they were exposed to suitable socioecological conditions and had evolved the capacity for open-ended vocal learning. This would solve the puzzle of why linguistic communication evolved in hominins, and not in other lineages that were facing similar social challenges (e.g. reliance on interdependent foraging in carnivores; cooperative breeding in callitrichid monkeys and meerkats).

Acknowledgements

We would like to thank all the collaborators, field assistants, students, and institutions involved in our orangutan communication project, but especially Maria van Noordwijk, Caroline Schuppli, and Uli Knief for their unwavering support over the years. M.F. was supported by a Freigeist Fellowship of the Volkswagen Foundation (grant 9B138). C.P.v.S. acknowledges the support of the NCCR Evolving Language Program (SNF #51NF40_180888).

References

Arnold, K. and Zuberbühler, K. (2012). 'Call combinations in monkeys: Compositional or idiomatic expressions?', *Brain and Language* 120(3): 303–309, doi:10.1016/j.bandl.2011.10.001.

Badihi, G., Graham, K. E., Fallon, B., et al. (2023). 'Dialects in leaf-clipping and other leaf-modifying gestures between neighbouring communities of East African chimpanzees', *Scientific Reports* 13(1): 147, doi:10.1038/s41598-022-25814-x.

Bakeman, R. and Quera, V. (2011). *Sequential Analysis and Observational Methods for the Behavioral Sciences.* Cambridge: Cambridge University Press.

9780192886491_Book.indb 515 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:04:59

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Bandini, E. and Harrison, R. A. (2020). 'Innovation in chimpanzees', *Biological Reviews* 95(5): 1167–1197, doi:10.1111/brv.12604.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2014). 'lme4: Linear mixed-effects models using Eigen and S4', *R Package Version* 1: 1–7, Retrieved from http://CRAN.R-project.org/package=lme4.

Boesch, C. and Boesch, H. (1989). 'Hunting behavior of wild chimpanzees in the Tai National Park', *American Journal of Physical Anthropology* 78: 547–573.

Boesch, C., Hohmann, G., and Marchant, L. F. (2002). *Behavioral Diversity in Chimpanzees and Bonobos*. Cambridge: Cambridge University Press.

Bolker, B. M., Brooks, M. E., Clark, C. J., et al. (2009). 'Generalized linear mixed models: A practical guide for ecology and evolution', *Trends in Ecology & Evolution* 24(3): 127–135, doi:10.1016/j.tree.2008.10.008.

Byrne, R. W., Cartmill, E., Genty, E., et al. (2017). 'Great ape gestures: Intentional communication with a rich set of innate signals', *Animal Cognition* 20(4): 755–769, doi:10.1007/s10071-017-1096-4.

Call, J. and Tomasello, M. (2007). *The Gestural Communication of Apes and Monkeys*. Mahwah, NY: Lawrence Erlbaum Associates.

Cartmill, E. and Byrne, R. (2010). 'Semantics of primate gestures: Intentional meanings of orangutan gestures', *Animal Cognition* 13(6): 793–804, doi:10.1007/s10071-010-0328-7.

Cartmill, E. A. (2023). 'Overcoming bias in the comparison of human language and animal communication', *Proceedings of the National Academy of Sciences* 120(47): e2218799120.

Cartmill, E. A. and Byrne, R. W. (2007). 'Orangutans modify their gestural signaling according to their audience's comprehension', *Current Biology* 17(15): 1345–1348, doi:10.1016/j.cub.2007.06.069.

Crockford, C. (2019). 'Why does the chimpanzee vocal repertoire remain poorly understood and what can be done about it?', in C. Boesch and R. Wittig (eds.), *The Chimpanzees of the Taï Forest: 40 Years of Research*, 394–409. Cambridge: Cambridge University Press.

Dingemanse, N. J. and Dochtermann, N. A. (2013). 'Quantifying individual variation in behaviour: Mixed-effect modelling approaches', *Journal of Animal Ecology* 82(1): 39–54, doi:10.1111/1365-2656.12013.

Dingemanse, N. J., Kazem, A. J., Réale, D., and Wright, J. (2010). 'Behavioural reaction norms: Animal personality meets individual plasticity', *Trends in Ecology & Evolution* 25(2): 81–89, doi:10.1016/j.tree.2009.07.013.

Dingemanse, N. J. and Wolf, M. (2013). 'Between-individual differences in behavioural plasticity within populations: Causes and consequences', *Animal Behaviour* 85(5): 1031–1039, doi:10.1016/j.anbehav.2012.12.032.

Doran, D. M. and Hunt, K. D. (1996). 'Comparative locomotor behavior of chimpanzees and bonobos', in R. W. Wrangham and W. C. McGrew (eds.), *Chimpanzee Cultures*, 93–108. Cambridge, MA: Harvard University Press.

Fichtel, C. and Van Schaik, C. P. (2006). 'Semantic differences in sifaka (Propithecus verreauxi) alarm calls: A reflection of genetic or cultural variants?', *Ethology* 112(9): 839–849.

Fischer, J. and Price, T. (2017). 'Meaning, intention, and inference in primate vocal communication', *Neuroscience & Biobehavioral Reviews* 82: 22–31, doi:10.1016/j.neubiorev.2016.10.014.

Fischer, J., Wheeler, B. C., and Higham, J. P. (2015). 'Is there any evidence for vocal learning in chimpanzee food calls?', *Current Biology* 25(21): R1028–R1029.

Fitch, W. T. (2010). *The Evolution of Language*. Cambridge: Cambridge University Press.

9780192886491_Book.indb 516 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:04:59

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Fitch, W. T. (2020). 'Animal cognition and the evolution of human language: Why we cannot focus solely on communication', *Philosophical Transactions of the Royal Society B: Biological Sciences* 375(1789): 20190046, doi:doi:10.1098/rstb.2019.0046.

Forss, S. I., Schuppli, C., Haiden, D., et al. (2015). 'Contrasting responses to novelty by wild and captive orangutans', *American Journal of Primatology* 77(10): 1109–1121.

Fox, E. A., van Schaik, C. P., Sitompul, A., and Wright, D. N. (2004). 'Intra- and interpopulational differences in orangutan (*Pongo pygmaeus*) activity and diet: Implications for the invention of tool use', *American Journal of Physical Anthropology* 125(2): 162–174.

Freeberg, T. M., Gentry, K. E., Sieving, K. E., and Lucas, J. R. (2019). 'On understanding the nature and evolution of social cognition: A need for the study of communication', *Animal Behaviour* 155: 279–286, doi:10.1016/j.anbehav.2019.04.014.

Fröhlich, M. and Hobaiter, C. (2018). 'The development of gestural communication in great apes', *Behavioral Ecology and Sociobiology* 72(12): 194, doi:10.1007/s00265-018-2619-y.

Fröhlich, M. and van Schaik, C. P. (2020). 'Must all signals be evolved? A proposal for a new classification of communicative acts', *Wiley Interdisciplinary Reviews: Cognitive Science* 11(4): e1527, doi:10.1002/wcs.1527.

Fröhlich, M. and van Schaik, C. P. (2022). 'Social tolerance and interactional opportunities as drivers of gestural redoings in orang-utans', *Philosophical Transactions of the Royal Society B: Biological Sciences* 377: 20210106.

Fröhlich, M., Wittig, R. M., and Pika, S. (2016a). 'Play-solicitation gestures in chimpanzees in the wild: Flexible adjustment to social circumstances and individual matrices', *Royal Society Open Science* 3(8): 160278, doi:10.1098/rsos.160278.

Fröhlich, M., Wittig, R. M., and Pika, S. (2016b). 'Should I stay or should I go? Initiation of joint travel in mother–infant dyads of two chimpanzee communities in the wild', *Animal Cognition* 19(3): 483–500, doi:10.1007/s10071-015-0948-z.

Fröhlich, M., Müller, G., Zeiträg, C., et al. (2017). 'Gestural development of chimpanzees in the wild: The impact of interactional experience', *Animal Behaviour* 134: 271–282, doi:10.1016/j.anbehav.2016.12.018.

Fröhlich, M., Kunz, J., Fryns, C., et al. (2020a). 'Social interactions and interaction partners in infant orang-utans of two wild populations', *Animal Behaviour* 166: 183–191, doi:10.1016/j.anbehav.2020.06.008.

Fröhlich, M., Müller, G., Zeiträg, C., et al. (2020b). 'Begging and social tolerance: Food solicitation tactics in young chimpanzees (*Pan troglodytes*) in the wild', *Evolution and Human Behavior* 41(2): 126–135, doi:10.1016/j.evolhumbehav.2019.11.002.

Fröhlich, M., Bartolotta, N., Fryns, C., et al. (2021a). 'Orangutans have larger gestural repertoires in captivity than in the wild—a case of weak innovation?', *iScience* 24(11): 103304, doi:10.1016/j.isci.2021.103304.

Fröhlich, M., Bartolotta, N., Fryns, C., et al. (2021b). 'Multicomponent and multisensory communicative acts in orang-utans may serve different functions', *Communications Biology* 4: 917, doi:10.1038/s42003-021-02429-y.

Fröhlich, M., van Schaik, C. P., van Noordwijk, M. A., and Knief, U. (2022). 'Individual variation and plasticity in the infant-directed communication of orang-utan mothers', *Proceedings of the Royal Society B: Biological Sciences* 289: 20220200.

Fröhlich, M., Noordwijk, M. A., Mitra Setia, T., et al. (2024). 'Wild and captive immature orangutans differ in their non-vocal communication with others, but not with their mothers', *Behavioral Ecology and Sociobiology* 78: 12, doi:10.1007/s00265-023-03426-3.

9780192886491_Book.indb 517 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:04:59

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Genty, E., Breuer, T., Hobaiter, C., and Byrne, R. W. (2009). 'Gestural communication of the gorilla (*Gorilla gorilla*): Repertoire, intentionality and possible origins', *Animal Cognition* 12(3): 527–546, doi:10.1007/s10071-009-0213-4.

Genty, E., Neumann, C., and Zuberbühler, K. (2015). 'Bonobos modify communication signals according to recipient familiarity', *Scientific Reports* 5: 16442.

Graham, K. E., Furuichi, T., and Byrne, R. W. (2020). 'Context, not sequence order, affects the meaning of bonobo (*Pan paniscus*) gestures', *Gesture* 19(2/3): 336–365.

Graham, K. E., Badihi, G., Safryghin, A., et al. (2022). 'A socio-ecological perspective on the gestural communication of great ape species, individuals, and social units', *Ethology Ecology & Evolution* 34(3): 235–259.

Grund, C., Badihi, G., Graham, K. E., et al. (2023). 'GesturalOrigins: A bottom-up framework for establishing systematic gesture data across ape species', *Behavior Research Methods*: doi:10.3758/s13428-023-02082-9

Halina, M., Rossano, F., and Tomasello, M. (2013). 'The ontogenetic ritualization of bonobo gestures', *Animal Cognition* 16(4): 653–666, doi:10.1007/s10071-013-0601-7.

Hardus, M. E., Lameira, A. R., Van Schaik, C. P., and Wich, S. A. (2009). 'Tool use in wild orang-utans modifies sound production: A functionally deceptive innovation?', *Proceedings of the Royal Society of London, Series B: Biological Sciences* 276(1673): 3689–3694, doi:10.1098/rspb.2009.1027.

Haslam, M. (2013). '"Captivity bias" in animal tool use and its implications for the evolution of hominin technology', *Philosophical Transactions of the Royal Society B: Biological Sciences* 368(1630): 20120421.

Hauser, M. D., Chomsky, N., and Fitch, T. W. (2002). 'The language faculty: What is it, who has it, and how did it evolve?', *Science* 298: 1568–1579.

Heesen, R., Bangerter, A., Zuberbühler, K., et al. (2020). 'Bonobos engage in joint commitment', *Science Advances* 6(51): eabd1306, doi:10.1126/sciadv.abd1306.

Hobaiter, C. and Byrne, R. W. (2011). 'The gestural repertoire of the wild chimpanzee', *Animal Cognition* 14(5): 747–767, doi:10.1007/s10071-011-0409-2.

Hobaiter, C. and Byrne, R. (2014). 'The meanings of chimpanzee gestures', *Current Biology* 24(14): 1596–1600, doi:10.1016/j.cub.2014.05.066.

Hockett, C. F. (1960). 'The origin of speech', *Scientific American* 203(3): 88–97.

Hopkins, W. D., Taglialatela, J. D., and Leavens, D. (2007). 'Chimpanzees differentially produce novel vocalizations to capture the attention of a human', *Animal Behaviour* 73(2): 281–286, doi:10.1016/j.anbehav.2006.08.004.

Knox, A., Markx, J., How, E., et al. (2019). 'Gesture use in communication between mothers and offspring in wild orang-utans (*Pongo pygmaeus wurmbii*) from the Sabangau Peat-Swamp Forest, Borneo', *International Journal of Primatology* 40(3): 393–416, doi:10.1007/s10764-019-00095-w.

Koolhaas, J. M., Korte, S. M., De Boer, S. F., et al. (1999). 'Coping styles in animals: Current status in behavior and stress-physiology', *Neuroscience & Biobehavioral Reviews* 23(7): 925–935, doi:10.1016/S0149-7634(99)00026-3.

Kopp, K. S. and Liebal, K. (2018). 'Conflict resolution in socially housed Sumatran orangutans (*Pongo abelii*)', *PeerJ* 6: e5303, doi:10.7717/peerj.5303.

Kummer, H. and Goodall, J. (1985). 'Conditions of innovative behaviour in primates', *Philosophical Transactions of the Royal Society B: Biological Sciences* 308(1135): 203–214, doi:10.1098/rstb.1985.0020.

Lameira, A. R., Hardus, M. E., Mielke, A., et al. (2016). 'Vocal fold control beyond the species-specific repertoire in an orang-utan', *Scientific Reports* 6: 30315, doi:10.1038/srep30315.

9780192886491_Book.indb 518 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:04:59

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Leca, J.-B., Gunst, N., and Huffman, M. A. (2010). 'Indirect social influence in the maintenance of the stone-handling tradition in Japanese macaques, Macaca fuscata', *Animal Behaviour* 79(1): 117–126, doi:10.1016/j.anbehav.2009.09.035.

Lehner, S. R., Burkart, J. M., and van Schaik, C. P. (2010). 'An evaluation of the geographic method for recognizing innovations in nature, using zoo orangutans', *Primates* 51(2): 101–118, doi:10.1007/s10329-009-0184-8.

Lemoine, S., Boesch, C., Preis, A., et al. (2020). 'Group dominance increases territory size and reduces neighbour pressure in wild chimpanzees', *Royal Society Open Science* 7(5): 200577, doi:doi:10.1098/rsos.200577.

Levinson, S. C. and Holler, J. (2014). 'The origin of human multi-modal communication', *Philosophical Transactions of the Royal Society B: Biological Sciences* 369(1651): 20130302, doi:10.1098/rstb.2013.0302.

Lewton, K. L. (2017). 'The effects of captive versus wild rearing environments on long bone articular surfaces in common chimpanzees (Pan troglodytes)', *PeerJ* 5: e3668, doi:10.7717/peerj.3668.

Liebal, K., Pika, S., and Tomasello, M. (2006). 'Gestural communication of orangutans (*Pongo pygmaeus*)', *Gesture* 6(1): 1–38, doi:10.1075/gest.6.1.02lie.

Liebal, K., Waller, B. M., Burrows, A. M., and Slocombe, K. E. (2013). *Primate Communication: A Multimodal Approach*. Cambridge: Cambridge University Press.

MacKinnon, J. (1974). 'The behaviour and ecology of wild orang-utans (*Pongo pygmaeus*)', *Animal Behaviour* 22(1): 3–74, doi:10.1016/s0003-3472(74)80054-0.

Manrique, H. M., Völter, C. J., and Call, J. (2013). 'Repeated innovation in great apes', *Animal Behaviour* 85(1): 195–202.

Maple, T. L. (1980). *Orangutan Behavior*. New York: Van Nostrand Reinhold.

Marler, P. (1965). 'Communication in monkeys and apes', in I. DeVore (ed.), *Primate Behaviour. Field Studies of Monkeys and Apes*, 544–584. New York: Holt, Rinehart and Winston.

Marler, P. (1967). 'Animal communication signals', *Science* 157(3790): 769–774, doi:doi:10.1126/science.157.3790.769.

Maynard Smith, J. and Harper, D. (2003). *Animal Signals*. Oxford: Oxford University Press.

Meulman, E. J., Sanz, C. M., Visalberghi, E., and van Schaik, C. P. (2012). 'The role of terrestriality in promoting primate technology', *Evolutionary Anthropology: Issues, News, and Reviews* 21(2): 58–68.

Mitani, J. C. and Watts, D. P. (2005). 'Correlates of territorial boundary patrol behaviour in wild chimpanzees', *Animal Behaviour* 70(5): 1079–1086.

Mitra Setia, T., Delgado, R., Utami Atmoko, S., et al. (2009). 'Social organization and male-female relationships', in S. A. Wich, S. Utami Atmoko, T. Mitra Setia, and C. P. van Schaik (eds.), *Orangutans: Geographic Variation in Behavioral Ecology and Conservation*, 245–253. Oxford: Oxford University Press.

Molesti, S., Meguerditchian, A., and Bourjade, M. (2020). 'Gestural communication in olive baboons (*Papio anubis*): Repertoire and intentionality', *Animal Cognition* 23(1): 19–40, doi:10.1007/s10071-019-01312-y.

Motes-Rodrigo, A. and Tennie, C. (2022). 'Captive great apes tend to innovate simple tool behaviors quickly', *American Journal of Primatology* 84(10): e23311.

Pika, S. and Fröhlich, M. (2019). 'Gestural acquisition in great apes: The Social Negotiation Hypothesis', *Animal Cognition* 22(4): 551–565, doi:10.1007/s10071-017-1159-6.

Ramsey, G., Bastian, M. L., and Van Schaik, C. (2007). 'Animal innovation defined and operationalized', *Behavioral and Brain Sciences* 30(4): 393–407, doi:10.1017/S0140525X07002373.

9780192886491_Book.indb 519 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:04:59

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Reader, S. M. and Laland, K. N. (2003). *Animal Innovation*. New York: Oxford University Press.
Rendall, D., Owren, M. J., and Ryan, M. J. (2009). 'What do animal signals mean?', *Animal Behaviour* 78(2): 233–240, doi:10.1016/j.anbehav.2009.06.007.
Rijksen, H. D. (1978). *A Fieldstudy on Sumatran Orang-utans (Pongo pygmaeus abelii, Lesson 1827): Ecology, Behaviour and Conservation*. Amsterdam: H. Veenman.
Robbins, M. M., Bermejo, M., Cipolletta, C., et al. (2004). 'Social structure and life-history patterns in western gorillas (*Gorilla gorilla gorilla*)', *American Journal of Primatology* 64(2): 145–159, doi:10.1002/ajp.20069.
Ruxton, G. D. and Schaefer, H. M. (2011). 'Resolving current disagreements and ambiguities in the terminology of animal communication', *Journal of Evolutionary Biology* 24(12): 2574–2585, doi:10.1111/j.1420-9101.2011.02386.x.
Schel, A. M., Bono, A., Aychet, J., et al. (2022). 'Intentional gestural communication amongst red-capped mangabeys (*Cercocebus torquatus*)', *Animal Cognition* 25(5): 1313–1330, doi:10.1007/s10071-022-01615-7.
Scott-Phillips, T. C. (2008). 'Defining biological communication', *Journal of Evolutionary Biology* 21(2): 387–395.
Seyfarth, R. M. and Cheney, D. L. (2014). 'The evolution of language from social cognition', *Current Opinion in Neurobiology* 28: 5–9, doi:10.1016/j.conb.2014.04.003.
Taborsky, B. and Oliveira, R. F. (2012). 'Social competence: An evolutionary approach', *Trends in Ecology & Evolution* 27(12): 679–688, doi:10.1016/j.tree.2012.09.003.
Taglialatela, J. P., Reamer, L., Schapiro, S. J., and Hopkins, W. D. (2012). 'Social learning of a communicative signal in captive chimpanzees', *Biology Letters* 8(4): 498–501, doi:doi:10.1098/rsbl.2012.0113.
Tennie, C., Bandini, E., Van Schaik, C. P., and Hopper, L. M. (2020). 'The zone of latent solutions and its relevance to understanding ape cultures', *Biology & Philosophy* 35(5): 1–42.
Tkaczynski, P. J., Mielke, A., Samuni, L., et al. (2020). 'Long-term repeatability in social behaviour suggests stable social phenotypes in wild chimpanzees', *Royal Society Open Science* 7(8): 200454, doi:10.1098/rsos.200454.
Tomasello, M. (1990). 'Cultural transmission in the tool use and communicatory signalling of chimpanzees', in S. T. Parker and K. R. Gibson (eds.), *'Language' and Intelligence in Monkeys and Apes: Comparative Developmental Perspectives*, 274–311. Cambridge: Cambridge University Press.
Tomasello, M. (2008). *Origins of Human Communication*. Cambridge, MA: MIT Press.
Tomasello, M., Call, J., Nagell, K., et al. (1994). 'The learning and use of gestural signals by young chimpanzees: A trans-generational study', *Primates* 35(2): 137–154, doi:10.1007/bf02382050.
van Schaik, C. P. (1999). 'The socioecology of fission-fusion sociality in orangutans', *Primates* 40(1): 69–86, doi:10.1007/BF02557703.
van Schaik, C. P. (2016). *The Primate Origins of Human Nature*. Hoboken, NJ: John Wiley & Sons.
van Schaik, C. P., Ancrenaz, M., Borgen, G., et al. (2003). 'Orangutan cultures and the evolution of material culture', *Science* 299: 102–105, doi:10.1126/science.1078004.
van Schaik, C. P., van Noordwijk, M. A., and Wich, S. A. (2006). 'Innovation in wild Bornean orangutans (*Pongo pygmaeus wurmbii*)', *Behaviour* 143(7): 839–876, doi:10.1163/156853906778017944.

9780192886491_Book.indb 520 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:04:59

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

van Schaik, C. P., Burkart, J., Damerius, L., et al. (2016). 'The reluctant innovator: Orangutans and the phylogeny of creativity', *Philosophical Transactions of the Royal Society B: Biological Sciences* 371(1690): 20150183.

Waller, B. M., Caeiro, C. C., and Davila-Ross, M. (2015). 'Orangutans modify facial displays depending on recipient attention', *PeerJ* 3: e827.

Watts, D. P. (2016). 'Production of grooming-associated sounds by chimpanzees (*Pan troglodytes*) at Ngogo: Variation, social learning, and possible functions', *Primates* 57(1): 61–72.

Whiten, A., Goodall, J., McGrew, W. C., et al. (1999). 'Cultures in chimpanzees', *Nature* 399(6737): 682–685, doi:citeulike-article-id:148764.

Wich, S. A., Swartz, K. B., Hardus, M. E., et al. (2009). 'A case of spontaneous acquisition of a human sound by an orangutan', *Primates* 50(1): 56–64, doi:10.1007/s10329-008-0117-y.

Wich, S. A., Vogel, E. R., Larsen, M. D., et al. (2011). 'Forest fruit production is higher on Sumatra than on Borneo', *PLoS ONE* 6(6): e21278, doi:10.1371/journal.pone.0021278.

Wich, S. A., Krützen, M., Lameira, A. R., et al. (2012). 'Call cultures in orang-utans?', *PLoS ONE* 7(5): e36180, doi:10.1371/journal.pone.0036180.

Zucker, E. L., Mitchell, G., and Maple, T. (1978). 'Adult male-offspring play interactions within a captive group of orang-utans (*Pongo pygmaeus*)', *Primates* 19(2): 379–384.

9780192886491_Book.indb 521 /12_first_proofs/first_pr[illegible]/xml_for_typesetting 27-Mar-25 23:04:59

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

9780192886491_Book.indb 522 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:04:59

CHAPTER 28

METHODS OF EXAMINING THE PROCESSING AND PRODUCTION OF RULE-BASED SEQUENCES IN ANIMALS

MÉLISSA BERTHET*, MAËL LEROUX*, SIMON W. TOWNSEND, AND STUART K. WATSON

28.1 Introduction

Language is a combinatorial system: meaningless sounds (phonemes) combine into meaningful units (words), which are likewise combined into larger, meaningful phrases. This process of combination is not random but governed by sets of conventionalized syntactic rules in each language (Berthet et al. 2023). These rules determine what combinations of units (e.g. combinations of sounds or words) are well formed, and which ones are not (Berthet et al. 2023). For example, our understanding of English syntax allows us to easily determine that the phrase 'man bites dog' is well formed while 'bites dog man' is not, despite being composed of an identical set of units. Being able to recognize, generalize, and apply these syntactic rules is fundamental to language use, and the question of when and how these capacities emerged in our lineage is, therefore, a crucial piece in the puzzle of language evolution.

Because the period of human history for which we have data on language use is extremely limited, shedding light on the evolution of syntactic processing abilities requires that we adopt a comparative approach, probing for similar capabilities in the

* These authors contributed equally.

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

cognition and communication of non-human animals (hereafter 'animals'). So far, this approach has been highly fruitful. Evidence of an ability to *process* complex rules within sequential structures appears to be present in non-human primates (hereafter 'primates') (Newport et al. 2004; Ravignani et al. 2013; Sonnweber et al. 2015; Watson et al. 2020), rodents (Toro and Trobalon 2005; de la Mora and Toro 2013), and birds (van Heijningen et al. 2009, 2013), indicating that some aspects of this ability may be taxonomically widespread, with ancient evolutionary origins. Studies of the communicative production abilities of animals have also found that some species (e.g. apes, monkeys, marine mammals, and birds; Engesser and Townsend 2019) are able to *produce* vocal sequences structured by combinatorial rules that bear similarities with some of those found in human language. In this chapter, we will review and critically evaluate the methods used to probe these capacities in animals. We begin by reviewing the methods used to probe the capacity to process abstract sequential rules. Specifically, we focus on 'artificial grammar experiments'—a highly productive and diverse set of experimental designs which have been applied broadly in animals and humans. In the second part of our chapter, we review the methods used in an important complementary research programme, which is the study of the rules underlying sequences spontaneously produced by animals and how they are interpreted.

28.2 Artificial grammar experiments: A powerful method of examining rule-processing abilities

'Artificial grammar experiments' are a popular and powerful family of experimental approaches for studying the ability of individuals (animal or human) to process rule-based structures. This umbrella term covers a diverse range of methodologies (Trotter et al. 2020), which can often be somewhat impenetrable for readers without a background in the field. Here, we attempt to unpack this topic by providing an introductory overview of the methods used in artificial grammar experiments. Rather than be prescriptive about how these studies should be designed, we aim to provide the reader with the foundational knowledge of salient methodological elements and their respective pros and cons, so that they can effectively evaluate and design such methods in future.

In broad terms, artificial grammar experiments consist of an exposure phase, followed by a test phase. In the exposure phase, subjects are repeatedly exposed to acoustic sequences that follow a predictable (i.e. rule-based) pattern. This exposure may occur passively, or as part of an active training task (we unpack these two approaches later in this chapter). In the test phase, the experiment measures the ability of subjects to i) recognize the rules underlying the pattern from the exposure phase, ii) generalize the rule to novel sequences, and iii) detect when the rule has been violated. To illustrate the rationale behind this design with a language-based analogy: demonstrating that one

9780192886491_Book.indb 524 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:04:59

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

understands the 'subject–verb–object' phrase structure that typifies English requires not only recognizing every instance of 'dog bites man' (a familiar sequence), but also readily understanding a never-before-heard construction comprised of the same categories of words, such as 'cat drinks water' (a generalized sequence), and finally that 'drinks water dog' is a violation of this structure. Identifying whether a subject can generalize a rule to novel items, and detect violations of it is, therefore, necessary to determine that they are truly processing the rule underlying the structure of the training sequences rather than simply rote memorizing them. A typical example of this design would be an experiment where subjects are passively habituated to repetitive streams of sequences, for example, 'A–B, A–B, A–B', from which subjects are expected to learn that sounds of category A are always followed by sounds of category B. This would be followed by a test phase in which the subject is probed with familiar sequences, sequences where the grammar is generalized ('A*i*–B*i*', where *i* represents a novel variant of the sound category) and sequences where the grammar is violated ('A–C') (e.g. Watson et al. 2020). In this chapter, we will focus on acoustic artificial grammar experiments because this represents the vast majority of the field, but it is worth noting that these questions can also be addressed in the visual domain (Sonnweber et al. 2015; Heimbauer et al. 2018; Milne et al. 2018a, b).

However, a non-trivial challenge emerges in precisely how one can measure a subject's performance in these tasks. While one can ask adult humans whether and how they recognized the rule underlying the test sequences, this is not feasible for prelinguistic humans and non-human animals. Below, we discuss the variety of approaches used to circumvent these difficulties under three headings related to key design considerations: grammar construction, active vs. passive designs, and choice of stimuli.

28.2.1 Grammar construction

A natural first step in designing an artificial grammar experiment is to decide on the grammar that will be presented to subjects, the format of which will depend on the specific phenomena of interest to the researchers and requires careful consideration to avoid confounding factors.

Let us consider a typical example: a topic that has attracted great interest is the capacity for human and non-human animals to process so-called 'non-adjacent dependencies', where a relationship exists between two or more elements in a sequence, separated by a string of other elements. For instance, in the English phrase, 'The dog who chased the cat was black', we understand that it is the dog who is black, not the cat. There is, therefore, a relationship or 'dependency' between the elements 'The dog' and 'was black', despite the distance between them. This non-adjacent dependency can be represented with the form 'A–B–B–A', where A represents a pair of dependent phrases (pertaining to the colour of the dog) which are separated by B, a string of intervening elements (note that this formulation only describes *one* of the *multiple* dependencies that exist within the sentence).

9780192886491_Book.indb 525 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:04:59

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

The investigation of the processing of non-adjacent dependencies often involves recreating the abstract aforementioned form 'A–B–B–A', with A and B representing different categories of stimuli. If one were to expose subjects to this grammar and then probe them with the sequence 'A–C–D–A', and they categorize this probe as being similar in structure to the training stimuli, can one safely assume they have processed the non-adjacent dependency? Unfortunately not. Just because a structure can be *described* as following a particular rule does not mean it must necessarily be *processed* using this rule if others are available. In this example, the subject could also be using a different heuristic such as, 'the first element must occur twice in the sequence' (which does not involve processing a non-adjacent dependency) to seemingly correctly categorize sequences. A significant hurdle in artificial grammar design is therefore devising a structure which excludes alternative heuristics or rules which are not of interest to the researchers.

One way to control for the issue described in the above example would be probing subjects with an additional sequence, such as 'A–B–A–B', which follows the confounding heuristic, but not the test grammar, and examine how the subject categorizes the sequence. However, even better would be to design a grammar which is immune to as many alternative heuristics as possible in the first place. For example, we could redesign our original grammar such that the first element is predictive of, but not identical to, the final element: 'A–B–C–D', 'A–B–B–D', 'B–A–A–C', 'B–D–A–C', and so on. Then, if a subject is presented with the sequence 'A–X–Y–D', the 'first element must be repeated' heuristic cannot lead to a false positive. Ruling out every possible confound for an artificial grammar is clearly not a trivial task and requires very careful consideration.

28.2.2 Active and passive designs

An important distinction in artificial grammar methodologies is whether an experiment makes use of an 'active' or 'passive' experimental task. We will first broadly characterize each of these approaches, and then contrast their respective benefits and constraints.

An active task is one in which the subject is provided with an explicit reward for a trained behaviour, which can be used to determine whether they have processed the grammar of interest. For example, van Heijningen et al. (2013) exposed zebra finches to a so-called Go/No Go task. In this task, birds were trained to understand that pecking a sensor after hearing a 'Go' stimulus (which conforms to a particular rule) resulted in a food reward, whereas pecking the sensor after a 'No Go' stimulus (which does not conform to the rule) resulted in a punishment (the lights turned off for fifteen seconds). Once the birds were reliably able to do this with training stimuli, they could be probed with novel sequences of varying structure. By observing how the trained birds categorized these novel stimuli, the authors were able to unambiguously determine which rule the birds were using to process the sequences.

'Passive' experiments, on the other hand, detect the propensity for subjects to *spontaneously* (i.e. without training) learn a rule underlying the sequences presented to

9780192886491_Book.indb 526 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:04:59

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

them. In the training phase, rather than having a task in which they must reach a criterion of performance (as in active tasks), subjects are passively exposed to a pre-defined number of training stimuli. In the test phase, probe stimuli are played to the subjects (typically a mix of familiar training sequences as well as ones which generalize or violate the rule) and their spontaneous behavioural responses towards these sequences are measured. Methods of measurement vary, but common dependent variables include measuring the latency (Reber et al. 2019), duration (Watson et al. 2020), and/or frequency (Ravignani et al. 2013) of head-turns towards the speaker in response to familiar, rule-generalized, and rule-violated probe sequences. This approach takes advantage of the fact that, for many animals (including humans), a normal behavioural response to hearing an unexpected acoustic stimulus is for the receiver to orient their gaze towards the source of the sound (Kemler Nelson et al. 1995; Gomez and Gerken 1999; Ravignani et al. 2013; Reber et al. 2019; Pougnault et al. 2020). We can therefore predict that if a subject processed the training grammar, then: (a) familiar sequences will elicit a very weak or no behavioural response, (b) sequences which generalize the rule will elicit the same or a slightly stronger response (the rule is familiar, but the constituent elements are novel), and (c) sequences which violate the rule will elicit the strongest response.

Active and passive designs each carry a number of important benefits and constraints. Active task experiments have the advantage that behavioural responses are clear and unambiguous to interpret: if an animal was trained to press a lever upon hearing a particular pattern and does so *only* when they hear this pattern, then one can be confident that lever pulls during the test phase indicate that some feature of the experimental stimulus directly elicited the trained association. Unfortunately, this reliability comes at the expense of extremely time-consuming training phases, potentially consisting of thousands of trials over many months. For example, Heimbauer et al. (2018) used a sample of two rhesus macaques (*Muccaca mulatta*) that required ~29,000 and ~50,000 trials to reach the training criterion in the training phase. These demands can make achieving a sample beyond a handful of individuals extremely difficult. However, the robustness of this approach also means that large subject samples may not be necessary to detect an experimental effect (Heimbauer et al. 2018; Jiang et al. 2018; Wakita 2019). Arguably, there are also some inferential limitations on such training-intensive paradigms. While it is certainly interesting that a given species can grasp complex sequential rules if given extensive training, it may be unclear what this work can tell us about the cognitive abilities and propensities of a 'species-typical' individual and their ecological relevance. This contrasts with passive designs, which test what subjects are able and motivated to learn without any extrinsic rewards. It is therefore important to carefully align the methods of a study with its specific set of research interests.

Another benefit of passive designs is that, in contrast to the time-consuming training involved in active task designs, the exposure phase can be relatively short. One can even carry it out in a group context, simultaneously habituating multiple individuals to the grammar, drastically reducing the time needed to carry out an experiment and minimizing the need for potentially stressful social separation. For example, Watson et al. (2020) habituated entire groups of common marmosets (*Callithrix jacchus*) and

9780192886491_Book.indb 527 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:05:00

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

chimpanzees (*Pan troglodytes*) to artificial grammar sequences for a total of five hours over one week while the groups went about their daily routine. Even this may have been overly cautious, as Reber et al. (2019) report that marmosets habituated to complex rules after just thirty minutes of familiarization. It is impossible to definitively say what a minimum duration would be for a passive exposure phase, but as far as we know there are no downsides to 'overexposure', so erring on the side of excess seems sensible. Passive exposure phases are also straightforward to implement, requiring relatively little time, equipment, or labour.

However, passive designs also lead to certain inferential challenges. Specifically, negative results can be very difficult to interpret: if a subject does not react to a test stimulus, how does one determine whether they were *unable* to process the underlying rules, or simply lacked the *motivation*? In our own study (Watson et al. 2020), for example, we found that five out of eight chimpanzees did not react at all during the testing phase of one of our experimental conditions. From this absence of evidence, one cannot infer that these individuals *could not* process the stimuli, nor even that they *did not* process it—perhaps they registered it on a cognitive level, but lacked the motivation to respond behaviourally. False positives are also a danger, as animals may turn their heads in the direction of the speaker for reasons unrelated to the experiment. It is, therefore, important to carefully consider what sample size and number of trials would be minimally necessary to convincingly statistically control for these types of inter- and intra-individual variation in behaviour: a response in a single trial may represent a false positive, but similar responses in five consecutive trials are likely to indicate a genuine response. Ongoing technological advances in biometrics for application to non-human species are one way these confounds could be overcome in future studies. For example, thermographic cameras (Brügger et al. 2021) and eye-trackers (Wilson et al. 2013, 2015) can detect fine-grained changes in arousal or attentional states, even in the absence of overt behavioural reactions. Unfortunately, these methods currently require prohibitively large investments of time and money relative to the resources available to many research groups.

28.2.3 Choice of stimuli

A final design consideration for artificial grammar experiments is what specific format the experimental stimuli should take. Many artificial grammar studies use artificial sounds, such as synthesized computer tones. Others have used sounds which, if not artificial, are at least largely foreign to the study species (e.g. human speech streams). The rationale behind using unfamiliar sounds is to mitigate any pre-existing associations that might shape subjects' behavioural responses when hearing the stimuli. An associated challenge is designing/choosing sounds which are sufficiently acoustically distinct from one another that the study species is readily able to categorize them differently. Pilot studies testing discriminability between the chosen stimuli for a study species are a sensible precaution. An interesting alternative is to use sounds drawn from

9780192886491_Book.indb 528 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:00

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

the vocal repertoire of the study species, ensuring that the subjects will be both able to discriminate between them and likely motivated to attend to them. Of course, care must be taken not to use sounds or combinations thereof that naturally elicit a strong behavioural response in receivers (e.g. alarm calls), but several studies have made effective use of song (Ouattara et al. 2009; van Heijningen et al. 2013; Chen and ten Cate 2017). An obvious benefit is also one of ecological relevance—observational studies have established that birdsong can be described as following a number of complex rules (see section 28.3), and artificial grammar experiments provide a means to test how the birds actually process them.

Although artificial grammar studies have produced impressive insights into the cognitive abilities of captive animals, the degree to which the specific capacities probed in many studies are ecologically relevant to the study subjects is unclear. Active tasks in particular require a large amount of training for individuals to learn the rules, indicating that although these capacities are within the absolute limits of cognition for a species, they cannot be readily accessed. Passive paradigms shore up these doubts somewhat, as individuals spontaneously learn the associations between elements, but are nevertheless carried out under highly artificial conditions with captive animals. Moreover, while the majority of the literature on this topic has examined perceptual processing abilities, a crucial topic from the perspective of understanding language evolution is how these abilities can be deployed in terms of signal production. In section 28.3, we will review the methods used to examine this.

28.3 Natural communication

The extent to which animal vocal systems are structured at the sequential level has received increasing attention over the last fifteen years (Ficken et al. 1978; Freeberg 2006; Ouattara et al. 2009; Candiotti et al. 2012; Jansen et al. 2012; Schlenker et al. 2016; Berthet et al. 2019). One particularly fruitful approach facilitating this work is the application of tools borrowed from mathematics and linguistics to animal systems to elucidate the rules governing the organization of natural vocal sequences (see review in Hurford 2011; Kershenbaum et al. 2016).

28.3.1 Identifying rules in the natural communicative sequences of animals

28.3.1.1 Identifying rule-based structures in animal song systems

Historically, the sequential dynamics of animal songs have received considerable research attention due to the highly variable nature of their structure (e.g. brown-tailed thrashers (*Toxostoma rufum*) produce around 1,000 song variants; Boughey and

9780192886491_Book.indb 529 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:05:00

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Thompson 1981). One popular approach to detecting rules underlying song structures involves calculating transition probabilities between song units. Markov chains are perhaps the most widely used method (Kershenbaum et al. 2016), whereby the probability of a unit's occurrence within a sequence is dependent on a finite number of previous units (i.e. 'first-order' Markov chains describe a unit that depends only on the previous unit, 'second-order' Markov chains depend on the two preceding units, etc.; see Figure 28.1). For example, if 'B' always follows 'A' in a sequence, regardless of what comes before 'A', then this rule can be captured by a first-order Markov chain. In contrast, a system such as hermit thrush (*Catharus guttatus*) song is best described by second-order Markov chains, suggesting that the occurrence of a unit depends on the two previously emitted units (McClean and Roach 2021).

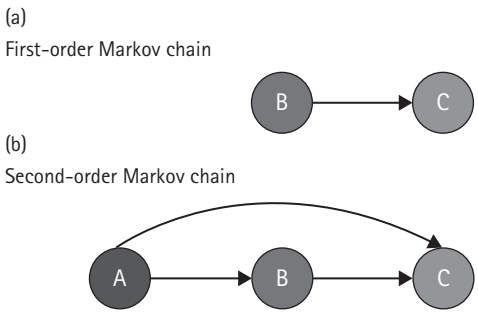


FIGURE 28.1 Schematic representations of a) a first-order Markov chain, where the occurrence of element C only depends on the occurrence of the previous element B, and b) a second-order Markov chain, where the occurrence of the element C depends on the occurrence of the previous elements AB.

The major limitation of Markov models is that a single model only provides the requested specific order-level of dependencies, rather than exploring whether (for example) first-, second-, or third-order provide the best fit for a given structure. To circumvent this issue, Markov chains can be implemented into 'prediction suffix trees'. These trees are visual representations of transition probabilities between units, constructed from a large set of Markov chains of increasing order and serve to detect which order of dependencies each unit's occurrence is best described with (Figure 28.2). For example, in canary songs, prediction suffix trees have demonstrated that the emission of a specific unit can be impacted by a unit emitted up to seven units prior (Markowitz et al. 2013).

Alternatively, recent advances in machine learning have identified complementary methods to detect longer dependencies with increased accuracy. For example, Morita et al. (2021) applied machine learning analyses to investigate rules governing the structures of Bengalese finch songs. Specifically, the authors measured the difference in the model's prediction accuracy when providing the model with full-length vs. truncated sequences, and found that feeding the model sequences of seven or fewer units significantly reduced its predictive accuracy, demonstrating at least an eight-order

9780192886491_Book.indb 530 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:00

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

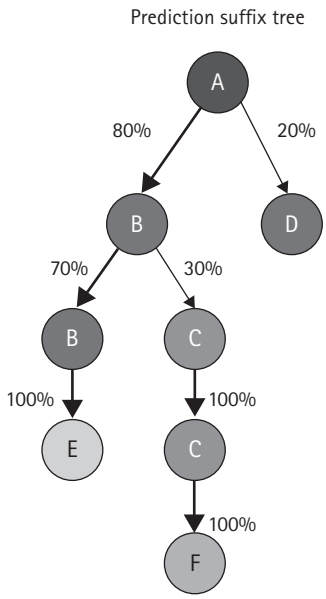


FIGURE 28.2 Schematic representations of a prediction suffix tree, where circles represent elements of a sequence, and arrows and their associated numbers represent the probability that an element follows another one.

Prediction suffix trees assess a wide range of Markov chains of varying orders, to determine the Markov chain that best describes the occurrence of each element. Here, the occurrence of F depends on the elements that were emitted up to four units prior (F occurs if the sequence ABCC has been emitted), while the probability of emission of D solely depends on the emission of element A.

dependency in this system. An advantage of machine learning-based approaches is that, unlike Markov chains, they do not make the *a priori* assumption that signal sequences are strictly governed by simple, linear, non-hierarchical rules. Indeed, whilst human language is deployed in linear sequences, its structure is represented hierarchically (Hauser et al. 2002), and recent evidence suggests this is the case for some animal songs too (Kershenbaum et al. 2014). However, machine learning algorithms require large amounts of data for their training, making their use empirically challenging for many systems (Katz 1987; Kneser and Ney 1995; Kershenbaum et al. 2016). Network analyses, however, whilst fundamentally Markovian in nature, represent one approach which can help elucidate the deep structure of a communication system and the relationship between its units with a reasonable amount of data. Specifically, network analyses help capture the properties of a combinatorial system through transition probabilities (i.e. similar to Markov chains) or, if these probabilities are unknown, via simple binary association. This makes network analyses widely applicable, especially when the system is extremely complex, with a large number of units precluding accurate estimation of transition probabilities (Boughey and Thompson 1981). For instance, Allen et al. (2019) applied such a method to humpback whale (*Megaptera novaeangliae*) songs and demonstrated they are composed of three to seven groups of highly connected units

9780192886491_Book.indb 531 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:05:00

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

(small-world networks). Network models highlight both the determinism and variety of transition patterns (or 'transition motifs') between elements (see Allen et al. 2019: fig. 1 for illustration) and have previously been used to quantify the presence of hierarchical structuring in, for example, real-world social networks (Chen and Fushing 2012). This may, therefore, represent a promising approach for detecting similar structuring in animal communication.

Another method to investigate non-Markovian relationships between units in song systems was introduced by Sainburg et al. (2019). These authors investigated hierarchical structuring in birdsong by analysing patterns of mutual information decay (i.e. the decreasing strength of the relationship between two units as the distance between them increases) in four species using three distinct models: a purely Markovian model, a purely hierarchical model (from Lin et al. 2017), and a combination of the two models using Markov chains as the end state of the hierarchical model. Contrary to Markov chains, a unit in a hierarchical model does not predict a single following unit but rather an array of new potential units. Furthermore, a hierarchical model is 'memoryless': a unit will, *independently of the preceding unit*, predict an array of new units. For Sainburg et al. (2019), the combined model provided the best fit for the data: sub-sequences of the songs were generated from a Markov process, composed of up to twenty-four dependencies, which were themselves organized on a higher-level hierarchy when distances became too large. Interestingly, this composite model is also the one that best describes human speech.

The extensive structural complexity displayed in the songs of many species represents a window into the evolution of combinatorial structuring. However, the comparison between songs and human language is limited by the lack of underlying semantics. Unlike compositional syntax in humans, where meaning-bearing units (i.e. words) are combined into larger structures (i.e. sentences; Berthet et al. 2023), the units composing songs seem to lack meaning (Berwick et al. 2011). If one is to investigate the evolutionary origins and drivers of human syntactic abilities, it is, therefore, paramount to investigate rules governing the combination of meaningful units into sequences in non-human animals. Contrary to song units, call types are usually meaning-bearing units (i.e. they occur in specific contexts and convey a specific set of information) (Berthet et al. 2023), and therefore represent a more promising parallel with the semantic layer of syntactic structuring in language. A further array of methods can be applied to these sequences to determine whether the manner of their combination augments or changes their meaning (described in Suzuki, Chapter 32, this volume), but the first step is to determine whether sequences of call types are produced in a rule-based manner.

28.3.1.2 Identifying rule-based structures in animal call systems

In the last two decades, the rule-governed nature of shorter, context-specific call sequences has received an upsurge in research interest (Zuberbuhler 2002; Ouattara et al. 2009; Jansen et al. 2012). However, until recently, researchers have refrained from quantifying the relationships between calls in sequences, instead deriving rules more qualitatively. One major consequence is that assumed relationships between elements

9780192886491_Book.indb 532 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:00

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

could be a mere product of happenstance juxtaposition, as opposed to non-random concatenation. For example, chimpanzees combine a 'pant-hoot' call (encoding identity) with a food call (denoting the presence of food) when conspecifics are joining a feeding party (Leroux et al. 2021). This combination could be the mere product of the chronological proximity of two call-eliciting contexts (individual joining + food), resulting in the simultaneous production of both calls sequentially. To demonstrate this combination follows specific rules of production, and therefore is a relevant structure in the species' communication system, one must show it is produced systematically rather than randomly.

To address this concern, researchers have applied some of the methods previously applied to birdsong to establish statistical relationships within call combinations. For instance, Girard-Buttoz et al. (2022) constructed a combinatorial network highlighting the transitional probabilities between the calls composing sequences observed in chimpanzees. In addition, the authors quantified whether units systematically occurred first or last in a sequence ('positional bias') and showed three calls (hoo, pant-grunt, and pant-hoot) occur in the first position of a sequence more often than expected by chance, while six typically occurred in the last position (pant-grunt, pant-bark, pant-scream, scream, bark, and pant).

Furthermore, a recent series of studies used a method developed in computational linguistics to identify non-random word combinations known as 'collocation analysis', and applied it to call combinations in various species (Leroux et al. 2021, 2022; Bosshard et al. 2022). This approach compares the relative co-occurrence of calls within the entire repertoire of observed call combinations, and statistically verifies which calls co-occur above chance level, distinguishing non-random association patterns from happenstance juxtapositions. For example, Leroux et al. (2022) implemented collocation analyses in the chimpanzee combinatorial vocal system and found that of the thirty-seven different types of combinations recorded, only fifteen were detected as non-random collocations. Although it is possible that some meaningful combinations are only produced very rarely, this approach suggests that more than half of the call combinations recorded in this study were most likely an incidental product of the random co-occurrence of calls. This data highlights the importance of the methods described here not only for identifying rule-based structures, but also for avoiding potential blind alleys caused by qualitative descriptions of call combinations. One key advantage of collocation analyses over other, more modelling-based approaches (e.g. Markov modelling), is that collocation analyses are easily applied and generalizable to various systems (see Bosshard et al. 2022 for an extensive step-by-step user guide). If we want to make meaningful comparisons across and within species, it is critical to standardize the field and apply, wherever possible, identical methodologies. Collocation analyses could represent one avenue to systematically capture combinatorial variability in animal call systems.

However, the identification of statistical regularities in animal call sequences does not necessarily entail that these are detectable by the animals themselves. In the next part, we present popular methods used to investigate species' capacity for rule processing of conspecific vocal sequences.

9780192886491_Book.indb 533 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:05:00

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

28.3.2 Receiver-based rule processing of natural animal sequences

Verifying whether the rules we have identified in the production of call combinations are processed by receivers can be determined by observing how they behave in response to sequences that follow the rule, vs. ones that do not. Two paradigms are commonly used to test this: simple stimulus presentation and habituation–discrimination experiments (see practical introductions in Zuberbühler and Wittig 2011; Fischer et al. 2013). In a simple stimulus presentation, subjects are presented with one stimulus type per trial, and their reaction is then compared between trials: if they react differently between trials, then subjects are assumed to perceive the difference between stimuli. For example, one could investigate ordering rule processing in the aforementioned Pant-Hoot + Food-Call combination in chimpanzees (Leroux et al. 2021) by conducting two distinct trials on separate occasions: one in which natural Pant-Hoot + Food-Call sequences are played back, and another playing back 'reverse-ordered' Food-call + Pant-hoot sequences. If individuals detect the rule, they should react differently to Pant-Hoot + Food-Call and Food-Call + Pant-Hoot sequences (e.g. approach vs. no reaction). In the habituation–dishabituation paradigm, subjects are presented with repeated occurrences of the same stimuli until they stop responding (habituation phase) and then immediately probed with a different stimulus (dishabituation phase): if they respond differently to the second stimulus, we can infer that they perceive the two stimuli differently. Applied to the example above, individuals would be presented with Pant-Hoot + Food-Call sequences until they stop responding, followed by a Food-Call + Pant-Hoot sequence: if they are able to detect the structural difference between the two sequences, they should respond with surprise at the probe sequence.

Two main types of stimuli sets can be used in these experiments. First, playbacks of a set of natural sequences that follow a given rule and other natural variations of the sequences can be used (e.g. Weiss et al. 2014; Andrieu et al. 2020). For example, Berthet et al. (2019) showed that, in titi monkeys' alarm sequence, the probability of transition between A- and B-calls depended on the type of predator encountered and its location. To investigate whether receivers process this rule, the authors presented titi monkeys with playbacks of natural sequences exhibiting varying probabilities of transitions. These playbacks elicited different gaze responses from the receivers, suggesting that the transition rules were perceived by the animals. The second typical stimuli set includes a combination of natural sequences with artificially modified sequences where the original rule has been violated (e.g. Briefer et al. 2013; Suzuki et al. 2016). For example, in a study investigating order effects in Japanese tits' combinatorial structures, subjects were presented with both a natural '*alarm call* + *recruitment call*' sequence and an artificially reversed '*recruitment call* + *alarm call*' sequence (Suzuki et al. 2016). The birds responded only to the sequences following the natural ordering rule, indicating that the rule of combination is salient to the birds. Together, these approaches have been extremely fruitful, identifying evidence of receiver processing of rule-based signals in species including birds (Suzuki et al. 2016, 2017; Engesser et al. 2016), mongooses

9780192886491_Book.indb 534 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:00

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

(Collier et al. 2020), monkeys (Arnold and Zuberbuhler 2006; Berthet et al. 2019), and apes (Clay and Zuberbühler 2011; Leroux et al. 2023).

28.4 Conclusion

In this chapter, we have highlighted the highly complementary nature of three methodologically diverse research programmes: artificial grammar experiments, the detection of combinatorial rules in animal signals, and the study of how animals process those signals. Each of these programmes employs a broad suite of specialized methodologies to explore different dimensions of the cognitive abilities that may represent the evolutionary building blocks of syntax. It is only with an accumulating body of data and evidence across a wide range of species and systems that we will be able to probe the extent to which animal communicative structuring informs us about the evolution of language and syntax.

Funding statement

The Swiss National Science Foundation partially or entirely funded the work of Melissa Berthet, Simon Townsend (Grant IDs: PP00P3_198912 and 315130_192620) and Stuart Watson (#51NF40_180888). Mael Leroux was supported by the French National Agency for Research and Rennes Métropole (Grant IDs: ANR-24-CE28-6179-01 and 24CO981).

References

Allen, J. A., Garland, E. C., Dunlop, R. A., and Noad, M. J. (2019). 'Network analysis reveals underlying syntactic features in a vocally learnt mammalian display, humpback whale song', *Proceedings of the Royal Society B: Biological Sciences* 286: 20192014, 10.1098/rspb.2019.2014.

Andrieu, J., Penny, S. G., Bouchet, H., et al. (2020). 'White-handed gibbons discriminate context-specific song compositions', *PeerJ* 8: e9477, 10.7717/peerj.9477.

Arnold, K. and Zuberbühler, K. (2006). 'Semantic combinations in primate calls', *Nature* 441: 303.

Berthet, M., Mesbahi, G., Pajot, A., et al. (2019). 'Titi monkeys combine alarm calls to create probabilistic meaning', *Science Advances* 5: eaav3991, 10.1126/sciadv.aav3991.

Berthet, M., Coye, C., Dezecache, G., and Kuhn, J. (2023). 'Animal linguistics: A primer', *Biological Reviews* 98(1): 81–98.

Berwick, R. C., Okanoya, K., Beckers, G. J. L., and Bolhuis, J. J. (2011). 'Songs to syntax: The linguistics of birdsong', *Trends in Cognitive Sciences* 15: 113–121, 10.1016/j.tics.2011.01.002.

Bosshard, A. B., Leroux, M., Lester, N. A., et al. (2022). 'From collocations to callocations: Using linguistic methods to quantify animal call combinations', *Behavioral Ecology and Sociobiology* 76: 122, 10.1007/s00265-022-03224-3.

9780192886491_Book.indb 535 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:05:00

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Boughey, M. J. and Thompson, N. S. (1981). 'Song variety in the brown thrasher (*Toxostoma rufum*)', *Zeitschrift für Tierpsychologie* 56: 47–58, https://doi.org/10.1111/j.1439-0310.1981.tb01283.x.

Briefer, E. F., Rybak, F., and Aubin, T. (2013). 'Does true syntax or simple auditory object support the role of skylark song dialect?', *Animal Behaviour* 86: 1131–1137, 10.1016/j.anbehav.2013.09.019.

Brügger, R. K., Willems, E. P., and Burkart, J. M. (2021). 'Do marmosets understand others' conversations? A thermography approach', *Scientific Advances* 7: eabc8790, 10.1126/sciadv.abc8790.

Candiotti, A., Zuberbühler, K., and Lemasson, A. (2012). 'Context-related call combinations in female Diana monkeys', *Animal Cognition* 15: 327–339, 10.1007/s10071-011-0456-8.

Chen, C. and Fushing, H. (2012). 'Multiscale community geometry in a network and its application', *Physical Review E* 86: 041120, 10.1103/PhysRevE.86.041120.

Chen, J. and ten Cate, C. (2017). 'Bridging the gap: Learning of acoustic nonadjacent dependencies by a songbird', *Journal of Experimental Psychology: Animal Learning and Cognition* 43: 295.

Clay, Z. and Zuberbühler, K. (2011). 'Bonobos extract meaning from call sequences', *PLOS ONE* 6: e18786, 10.1371/journal.pone.0018786.

Collier, K., Radford, A. N., Stoll, S., et al. (2020). 'Dwarf mongoose alarm calls: Investigating a complex non-human animal call', *Proceedings of the Royal Society B Biological Sciences* 287: 20192514, 10.1098/rspb.2019.2514.

de la Mora, D. M. and Toro, J. M. (2013). 'Rule learning over consonants and vowels in a non-human animal', *Cognition* 126: 307–312, 10.1016/j.cognition.2012.09.015.

Engesser, S. and Townsend, S. W. (2019). 'Combinatoriality in the vocal systems of nonhuman animals', *WIREs Cognitive Science* 10: e1493, 10.1002/wcs.1493.

Engesser, S., Ridley, A. R., and Townsend, S. W. (2016). 'Meaningful call combinations and compositional processing in the southern pied babbler', *Proceedings of the National Academy of Sciences* 113: 5976–5981, 10.1073/pnas.1600970113.

Ficken, S. R., Millicent S., Ficken, R. W., and Witkin, S. R. (1978). 'Vocal repertoire of the black-capped chickadee', *The Auk* 95: 34–48.

Fischer, J., Noser, R., and Hammerschmidt, K. (2013). 'Bioacoustic field research: A primer to acoustic analyses and playback experiments with primates', *American Journal of Primatology* 75: 643–663.

Freeberg, T. M. (2006). 'Social complexity can drive vocal complexity: Group size influences vocal information in Carolina chickadees', *Psychological Science* 17: 557–561, 10.1111/j.1467-9280.2006.01743.x.

Girard-Buttoz, C., Zaccarella, E., Bortolato, T., et al. (2022). 'Chimpanzees produce diverse vocal sequences with ordered and recombinatorial properties', *Communications Biology* 5: 1–15, 10.1038/s42003-022-03350-8.

Gomez, R. L. and Gerken, L. (1999). 'Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge', *Cognition* 70: 109–135, 10.1016/S0010-0277(99)00003-7.

Hauser, M. D., Chomsky, N., and Fitch, W. T. (2002). 'The faculty of language: What is it, who has it, and how did it evolve?', *Science* 298: 1569–1579.

Heimbauer, L. A., Conway, C. M., Christiansen, M. H., et al. (2018). 'Visual artificial grammar learning by rhesus macaques (*Macaca mulatta*): Exploring the role of grammar complexity and sequence length', *Animal Cognition* 21: 267–284, 10.1007/s10071-018-1164-4.

Hurford, J. R. (2011). *The Origins of Grammar*. Oxford: Oxford University Press.

Jansen, M. B., David A. W. A. M., Cant, M. A., and Manser, M. B. (2012). 'Segmental concatenation of individual signatures and context cues in banded mongoose (*Mungos mungo*) close calls', *BMC Biology* 10: 1–11.

9780192886491_Book.indb 536 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:00

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Jiang, X., Long, T., Cao, W., et al. (2018). 'Production of supra-regular spatial sequences by macaque monkeys', *Current Biology* 28: 1851–1859.

Katz, S. (1987). 'Estimation of probabilities from sparse data for the language model component of a speech recognizer', *IEEE Transactions on Acoustics, Speech, and Signal Processing* 35: 400–401, 10.1109/TASSP.1987.1165125.

Kemler Nelson, D. G., Jusczyk, P. W., Mandel, D. R., et al. (1995). 'The head-turn preference procedure for testing auditory perception', *Infant Behavioural Development* 18: 111–116, 10.1016/0163-6383(95)90012-8.

Kershenbaum, A., Bowles, A. E., Freeberg, T. M., et al. (2014). 'Animal vocal sequences: Not the Markov chains we thought they were', *Proceedings of the Royal Society B: Biological Sciences* 281: 20141370, 10.1098/rspb.2014.1370.

Kershenbaum, A., Blumstein, D. T., Roch, M. A., et al. (2016). 'Acoustic sequences in non-human animals: A tutorial review and prospectus', *Biological Reviews* 91: 13–52, https://doi.org/10.1111/brv.12160.

Kneser, R. and Ney, H. (1995). 'Improved backing-off for M-gram language modeling', in *1995 International Conference on Acoustics, Speech, and Signal Processing*, 1: 181–184, 10.1109/ICASSP.1995.479394.

Leroux, M., Bosshard, A. B., Chandia, B., et al. (2021). 'Chimpanzees combine pant hoots with food calls into larger structures', *Animal Behaviour* 179: 41–50, 10.1016/j.anbehav.2021.06.026.

Leroux, M., Chandia, B., Bosshard, A. B., et al. (2022). 'Call combinations in chimpanzees: A social tool?', *Behavioural Ecology* 33(5): 1036–1043.

Leroux, M., Schel, A. M., Wilke, C., et al. (2023). 'Call combinations in wild chimpanzees: Evidence for compositional syntax?', *Nature Communications* 14: https://doi.org/10.1038/s41467-023-37816-y.

Lin, H. W., Tegmark, M., and Rolnick, D. (2017). 'Why does deep and cheap learning work so well?', *Journal of Statistical Physics* 168: 1223–1247, 10.1007/s10955-017-1836-5.

Markowitz, J. E., Ivie, E., Kligler, L., and Gardner, T. J. (2013). 'Long-range order in canary song', *PLOS Computational Biology* 9: e1003052, 10.1371/journal.pcbi.1003052.

McLean, L. C. and Roach, S. P. (2021). 'Markov dependencies in the song syntax of hermit thrush (*Catharus guttatus*)', *Journal of Ornithology* 162: 469–476, 10.1007/s10336-020-01840-2.

Milne, A. E., Petkov, C. I., and Wilson, B. (2018a). *Auditory and Visual Sequence Learning in Humans and Monkeys using an Artificial Grammar Learning Paradigm*. R Foundation for Statistical Computing.

Milne, A., Wilson, B., and Christiansen, M. (2018b). 'Structured sequence learning across sensory modalities in humans and nonhuman primates', *Current Opinions in Behavioural Science* 21: 39–48, 10.1016/j.cobeha.2017.11.016.

Morita, T., Koda, H., Okanoya, K., and Tachibana, R. O. (2021). 'Measuring context dependency in birdsong using artificial neural networks', *PLOS Computational Biology* 17: e1009707, 10.1371/journal.pcbi.1009707.

Newport, E. L., Hauser, M. D., Spaepen, G., and Aslin, R. N. (2004). 'Learning at a distance II. Statistical learning of non-adjacent dependencies in a non-human primate', *Cognitive Psychology* 49: 85–117, 10.1016/j.cogpsych.2003.12.002.

Ouattara, K., Lemasson, A., and Zuberbühler, K. (2009). 'Campbell's monkeys concatenate vocalizations into context-specific call sequences', *Proceedings of the National Academy of Sciences* 106: 22026–22031, 10.1073/pnas.0908118106.

Pougnault, L., Levréro, F., Mulot, B., and Lemasson, A. (2020). 'Breaking conversational rules matters to captive gorillas: A playback experiment', *Scientific Reports* 10: 6947, 10.1038/s41598-020-63923-7.

9780192886491_Book.indb 537 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:05:01

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Ravignani, A., Sonnweber, R.-S., Stobbe, N., and Fitch, W. T. (2013). 'Action at a distance: Dependency sensitivity in a New World primate', *Biology Letters* 9: 20130852.

Reber, S. A., Šlipogor, V., Oh, J., et al. (2019). 'Common marmosets are sensitive to simple dependencies at variable distances in an artificial grammar', *Evolution of Human Behaviour* 40: 214–221.

Sainburg, T., Theilman, B., Thielk, M., and Gentner, T. Q. (2019). 'Parallels in the sequential organization of birdsong and human speech', *Nature Communications* 10: 1–11, 10.1038/s41467-019-11605-y.

Schlenker, P., Chemla, E., Schel, A. M., et al. (2016). 'Formal monkey linguistics', *Theoretical Linguistics* 42: 1–90, 10.1515/tl-2016-0001.

Sonnweber, R., Ravignani, A., and Fitch, W. T. (2015). 'Non-adjacent visual dependency learning in chimpanzees', *Animal Cognition* 18: 733–745.

Suzuki, T. N., Wheatcroft, D., and Griesser, M. (2016). 'Experimental evidence for compositional syntax in bird calls', *Nature Communications* 7: 10986, 10.1038/ncomms10986.

Suzuki, T. N., Wheatcroft, D., and Griesser, M. (2017). 'Wild birds use an ordering rule to decode novel call sequences', *Current Biology* 27: 2331–2336.

Toro, J. M. and Trobalón, J. B. (2005). 'Statistical computations over a speech stream in a rodent', *Attention, Perception & Psychophysics* 67: 867–875, 10.3758/BF03193539.

Trotter, A. S., Monaghan, P., Beckers, G. J. L., and Christiansen, M. H. (2020). 'Exploring variation between artificial grammar learning experiments: Outlining a meta-analysis approach', *Topics in Cognitive Science* 12: 875–893, 10.1111/tops.12454.

van Heijningen, C. A., De Visser, J., Zuidema, W., and Ten Cate, C. (2009). 'Simple rules can explain discrimination of putative recursive syntactic structures by a songbird species', *Proceedings of the National Academy of Sciences* 106: 20538–20543.

van Heijningen, C. A., Chen, J., van Laatum, I., et al. (2013). 'Rule learning by zebra finches in an artificial grammar learning task: Which rule?', *Animal Cognition* 16: 165–175.

Wakita, M. (2019). 'Auditory sequence perception in common marmosets (Callithrix jacchus)', *Behavioural Processes* 162: 55–63, 10.1016/j.beproc.2019.01.014.

Watson, S. K., Burkart, J. M., Schapiro, S. J., et al. (2020). 'Nonadjacent dependency processing in monkeys, apes, and humans', *Scientific Advances* 6: 10.1126/sciadv.abb0725.

Weiss, M., Hultsch, H., Adam, I., et al. (2014). 'The use of network analysis to study complex animal communication systems: A study on nightingale song', *Proceedings of the Royal Society B: Biological Sciences* 281: 20140460, 10.1098/rspb.2014.0460.

Wilson, B., Slater, H., Kikuchi, Y., et al. (2013). 'Auditory artificial grammar learning in macaque and marmoset monkeys', *Journal of Neuroscience* 33: 18825–18835, 10.1523/JNEUROSCI.2414-13.2013.

Wilson, B., Smith, K., and Petkov, C. I. (2015). 'Mixed-complexity artificial grammar learning in humans and macaque monkeys: evaluating learning strategies', *European Journal of Neuroscience* 41: 568–578, 10.1111/ejn.12834.

Zuberbühler, K. (2002). 'A syntactic rule in forest monkey communication', *Animal Behaviour* 63: 293–299, 10.1006/anbe.2001.1914.

Zuberbühler, K. and Wittig, R. M. (2011). *Field Experiments with Non-human Primates: A Tutorial*. Cambridge: Cambridge University Press.

9780192886491_Book.indb 538 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:01

CHAPTER 29

BIOACOUSTICS AND RHYTHM

DARIA VALENTE AND ANDREA RAVIGNANI

29.1 Bioacoustics—areas of investigation

Human language is composed of multiple traits, and so is speech (Jarvis 2019). Some of these traits can be found ubiquitously in the animal realm, others can be shared with few species, or even be uniquely human (Fitch 2017). Universal components of speech are auditory and vocal usage learning. The first is the capability to acquire and memorize novel sound associations, the second refers to the emission of sounds (either innate or learned) in unusual circumstances (Petkov and Jarvis 2012). A whole different level is represented by the vocal production learning, or the ability to mimic sounds. This trait (commonly referred to as vocal learning) is only found in species belonging to two animal classes: mammals (humans, bats, elephants, cetaceans, pinnipeds) and birds (songbirds, parrots, hummingbirds). Bioacoustic research can provide critical insight into the evolutionary roots of acoustic communicative systems, spoken language included. This field of investigation mainly relies on comparative studies focused on non-human animal species sharing fundamental traits with humans. Within phonetics and speech sciences, we can opt for two broad spheres of action: we can either compare our vocal system to that of non-human primates, our closest living relatives, or focus on other species which acquired language-relevant traits via convergent evolution.

29.2 Recordings

Acoustic recordings can be extremely diverse. We should define both the species and the vocal sequence to be recorded, as well as the environment, e.g. marine or terrestrial, in which the animals live, and the medium—i.e. air vs. water—in which the

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

vocalization is emitted. Recording in natural vs. captive environments also entails further considerations. For instance, recording in a rainforest or in a zoo may entail sound attenuation or reverberation (see Table 29.1 for the explanation of these and other discipline-specific terms). Heavy wind can potentially impact recording quality of species in open habitats, as in the cases of geladas (*Theropithecus gelada*) foraging on Ethiopian cliffs (Gustison et al. 2016) or males of the northern elephant seal (*Mirounga angustirostris*) fighting for sexual competition on California beaches (Mathevon et al. 2017). These are only a few examples of the many features to account for when planning data collection.

When collecting data, environmental conditions should be factored in, together with the relationship between weather conditions and vocal behaviour. Indeed, across different taxa, environmental factors strongly influence vocal behaviour. For instance, temperature, presence, and amount of rain may influence the calling behaviour of kloss (*Hylobates klossii*), agile (*Hylobates albibarbis*), eastern hoolock (*Hoolock leuconedys*), and Mueller's Bornean gibbons (*Hylobates muelleri*) (Clink et al. 2020). Other sources of disturbance might inhibit vocalizations, result in lower-quality sound recordings, or both; these disturbances include other animals vocalizing, and non-biotic factors such as wind, waves, vessels, traffic, in both terrestrial and aquatic environments. Environmental conditions do not only affect the behaviour of animals. Extreme temperatures and humidity can oxidize recorders, microphones, and cables. Desiccants are highly recommended to minimize moisture issues. We recommend using waterproof gear when collecting data on species inhabiting tropical environments, such as rainforests (Browning et al. 2017). Likewise, proper maintenance of equipment used in marine environments requires an accurate rinse to prevent rust deposits and corrosion. Different environments spur different considerations. Obvious sources of noise in captivity (e.g. zoos) are visitors, animal keepers, and employees. To minimize their impact, researchers should privilege the time of day in which animals are most active while avoiding recording at crowded times. Perhaps less obvious, lights, air conditioners, and power generators also produce a constant noise. A good practice is conducting a preliminary assessment to identify potential quiet and noisy zones within the enclosures (Browning et al. 2017). Indeed, although modern exhibits are built to minimize these phenomena, the floor and walls themselves represent a surface source of reverberations and echoes (Dilungana et al. 2022).

From a behavioural perspective, we should be aware of both diel and seasonal variations of the vocal pattern of a species. In addition, vocal displays are not fixed and immutable but, rather, likely to change across contexts (Torti et al. 2013) and ages; for instance, temporal features of seals and indris show variations during ontogeny (De Gregorio et al. 2021a; Ravignani et al. 2019b). We should also keep in mind that the environment in which an individual vocalizes can change throughout the life cycle of that individual; indeed, while adult males of *Phoca vitulina* mostly sing underwater, pups' lives are at first closely tied to land (Riedman 1990; Duengen and Ravignani 2023). All this has, of course, repercussions on the sounds produced and how they should be best recorded.

9780192886491_Book.indb 540 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:01

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Table 29.1 Definitions and explanations of common words (mostly stemming from biology or physical acoustics) used in bioacoustics, listed alphabetically.

Active recording	Direct data collection operated by human observers
array of microphones (i.e. acoustical array)	A set (between two and hundreds) of microphones that can capture sound from multiple directions, used for large-scale monitoring systems.
attenuation	The loss of energy of a sound while propagating in a medium.
diel	Describing a process which repeats with a 24h cycle.
DTAG	Digital Acoustic Recording Tag, a non-invasive device developed for monitoring marine mammals' behaviour. A tag usually includes a large array of sensors and can record continuously through a built-in hydrophone.
dynamic range	Difference between the highest and lowest sound pressure levels that can be recorded.
echo	Sound reflected (or returned) with sufficient magnitude and delay to be detectable as a distinct sound from the originally transmitted one.
frequency	Number of repetitions of an event per unit of time.
frequency response	Frequency-specific output sensitivity of a microphone. It details the range of frequencies that can be captured/recorded.
high-pass filter (i.e. bass-cut filter)	Transformation of the audio signal that only retains frequencies above a given threshold, and 'erases' all sounds with a frequency below that threshold.
hydrophone	Microphone specifically designed for underwater recording.
infrasonic	Sound below the human audible range, that is, below 20 Hz.
noise	Unwanted sound which can decrease the quality, clarity, and ability to properly analyse a target sound. Above some thresholds, noise can be physically harmful.
Nyquist frequency	In digital recordings, it quantifies the highest frequency which can be sampled from a signal while avoiding aliasing (i.e. a type of distortion), Quantitatively, the Nyquist frequency is half of the digital sampling frequency.
passive recording	Collecting data on wildlife and environment using autonomous recorders, usually acoustic sensors or camera traps, only sporadically operated by humans.
proximity effect	Exaggeration of low-frequency signals in a directional microphone when the microphone is located near source.
reverberation	Persistence of a sound after the source has stopped, caused by multiple reflection or scattering of the sound within a closed environment.
sampling frequency	In digital recordings, it quantifies the number of discrete samples in time of a waveform; in other words, a signal sampled at 48,000 Hz is described by 48,000 snapshots, each second, of the position of the wave.
signal to noise ratio	Ratio between the power of a signal and the power of noise.

(continued)

9780192886491_Book.indb 541 /12_first_proofs/first_pr[illegible]/xml_for_typesetting 27-Mar-25 23:05:01

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Table 29.1 Continued

Active recording	Direct data collection operated by human observers
singing	Able to emit songs, series of notes of different types characterized by a frequency variation, uttered following a hierarchical structure. Singing may occur with different degrees of interaction with conspecifics. In solos only a singer is involved. Duets involve a dyadic interaction. Choruses include multiple singers.
sound pressure	The difference in pressure caused by a sound wave. Microphones and sound pressure level meters can measure sound pressure.
soundtrap	Underwater sound recorder; a particular sort of hydrophone.
ultrasonic	Sound above the human audible range, that is, above 20,000 Hz.
vocal pattern	Vocal behaviour of a species, also in relation to the time of day or year in which the species habitually vocalizes. Vocalizations can include a sequence of calls or notes emitted by a non-human animal which can form a song or another vocal sequence.
waveform	Function and/or graph of a signal describing its amplitude (y-axis) over time (x-axis).

Data collection for bioacoustics relevant to language evolution typically relies on direct (i.e. active) recordings. By contrast, passive acoustics is an established methodology for soundscape ecology, ecological research (Melo et al. 2021), or for monitoring a species in relation to its ecological requirements (e.g. to evaluate species richness, abundance, occupancy models, vocal behaviour throughout the day; Ravaglia et al. 2023). Nonetheless, the use of non-invasive passive recorders could be useful for monitoring animal communities at a larger scale, especially for collecting data on elusive species living in dense or aquatic environments, that are crucial for investigating the evolution of vocal behaviour in non-human animals, e.g. singing primates or pinnipeds. These species have indeed been involved in studies investigating common patterns defining the evolution of vocal behaviour across species (De Gregorio et al. 2021b; Duengen and Ravignani 2022; Raimondi et al. 2023). Recording these species can be challenging, especially when it comes to recording vocalizations emitted underwater (Ravignani et al. 2019a). In these contexts, the identification of the vocalizing individual is often infeasible because of the recording setup, usually consisting of arrays of microphones recording passively. Nevertheless, passive recordings can still provide indirect evidence of rhythmicity (Ravignani et al. 2019a); they are particularly useful for those species canonically associated with comparative studies on language evolution vocalizing outside the human observer audible range, like bats and elephants (see definitions of infrasonic and ultrasonic in Table 29.1), or living in peculiar environments, like rainforests and oceans.

Once the target environment, species, vocalization types, etc., are chosen, how does one actually go about to record their subjects? The first step involves acquiring a microphone and a sound recorder.

9780192886491_Book.indb 542 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:01

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

29.2.1 Devices

When choosing a microphone, its directionality is a key parameter to evaluate (Rhinehart et al. 2020). Broadly speaking, omnidirectional microphones are optimal for detecting sounds from all directions and therefore for recording soundscapes; directional microphones are ideal when recording a specific individual in a particular direction, and for attenuating noise coming from directions other than the signal's one, for instance a close river. The advantage of using a directional microphone is improving the signal to noise ratio by reducing environmental sounds, important as much in the wild as in captivity, where it allows mitigating reverberations and echoes (Rhinehart et al. 2020). Among many kinds of directional microphones available, the use of shotgun highly directional microphones seems to be preferred. Condenser shotgun microphones, such as Sennheiser ME 66 and ME 67 (frequency response: 40–20,000 Hz) are recommended when microphones must be necessarily placed at a distance from the source. These have been used, among innumerable examples, to record pinnipeds in captivity (*Phoca vitulina*: Ravignani et al. 2019b; Torres Borda et al. 2021) as well as primates (*Indri indri*: Gamba et al. 2016; De Gregorio et al. 2019; *Theropithecus gelada*: Gustison et al. 2016), and other mammals (*Procavia capensis*: Demartsev et al. 2022) in the wild. Condenser shotgun microphones AKG C98 or Neumann KMR 82i (frequency response: 20–20,000 Hz) seems to be common to record animals in the wild (*Indri indri*: De Gregorio et al. 2019; Valente et al. 2021; *Mirounga angustirostris*: Mathevon et al. 2017). Other singing primate species (De Gregorio et al. 2022) have been recorded using different condenser microphones; Clink and Lau (2020) relied on Røde NT-USB and NTG-2, and Marantz Professional Audio Scope SG-5B to record tarsiers, titi monkeys, and gibbons (*Tarsius spectrumgurskyae*; *Plecturocebus cupreus*; *Hylobates funereus*) in the wild. Besides being excellent for recording sounds from a distance, high-quality directional microphones allow us to compensate for the proximity effect, enabling use at the closer ranges.

Shotgun microphones are based on the interference tube principle, with the microphone commonly placed at the end of the tube. The longer the tube and the higher the frequency of the recorded signals, the higher the directivity (sensitivity to signals): this means that when recording high frequencies (from 4,000 Hz), the receiving portion is rather narrow. At the same time, the directionality decreases as the incoming frequency decreases. By consequence, lower frequencies will be less attenuated than high ones (Madhusudhana et al. 2022). In other words, mispointing a microphone will negatively affect recording of high frequencies (e.g. bird chirps) significantly more than low ones (e.g. lion roars). High-quality microphones possess filters to cut off low frequencies; for instance, the AKG C98 has a built-in high-pass filter with onset at 100 Hz, the Neumann KMR 82i integrates a high-pass filter suppressing sources of interference such as wind and handling noise. This feature is ideal when recording high frequencies and in noisy conditions (Madhusudhana et al. 2022). Choosing proper settings has important ramifications. For instance, Ehnes and Foote (2015) found that directional microphones

9780192886491_Book.indb 543 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:05:01

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

(Sennheiser MKH-70) showed better performance than omnidirectional ones (Wildlife Acoustics' SMX-II weatherproof microphone combined with the SM2 + Terrestrial Acoustic Package) in discriminating songs of ovenbirds (*Seiurus aurocapillus*).

The scenario changes when dealing with underwater or ultrasonic recordings. In recent years, sperm whales (*Physeter macrocephalus*) have been sampled through DTAGs providing stereo recordings with a flat frequency response of 0.5–50 kHz, custom-built hydrophones (Benthos AQ-4 elements with a flat frequency response of 0.1–30 kHz), or hydrophone soundtraps (SoundTrap 300 with a flat frequency response of 20–60,000 Hz). Other studies relied on HTI-96 hydrophone with a frequency response of 2–30 kHz (*Physeter macrocephalus*: Tønnesen et al. 2018). For recording terrestrial species vocalizing in the ultrasonic range, one of the most efficient devices is the ultrasonic microphone Avisoft USG 116Hme with condenser microphone CM16, a recorder with a frequency response of 20 Hz–460 kHz. For example, this has been used by Knörnschild et al. (2013) to record calls of *Carollia perspicillata*.

At this point, we should identify sampling rates and frequency appropriate not only to the vocalizing species, but also to a given vocalization. When converting analogue to digital signals, one needs to balance the maximum recordable frequency with the sampling frequency. The golden standard in the field seems to be a **sampling frequency between** 44.1 and 48 kHz (comparable to that of music CDs). However, when deciding these settings, one should mind the frequency range of the target species: given that the frequency is a function of the **Nyquist frequency** (half of the sampling frequency), choosing a frequency value too close to the Nyquist frequency might be prone to artefacts such as aliasing (Shannon 1949). In other words, the sampling frequency should be at least twice the maximum frequency of the sound one wants to reliably record. In bioacoustic research, the most common **resolution** is between 16 and 24 bits, allowing a **dynamic range** of about 96 and 144 dB, respectively. These requirements can be fulfilled by using most solid-state digital audio recorders available. Common producers are Marantz (PMD-222, PMD-660, PMD-661), Tascam (HD-P2, DR-100 MKII, DR-40, and DR-05), Zoom (H5, F8), Sound Devices (702), and Olympus (S100 and LS05). When recording a vocal display, a researcher might be tempted to use a video camera to simultaneously record also the focal animal behaviour. While, in principle, one could decide to simplify the process by using just the camera, the latter has microphones with inadequate frequency responses, impairing the quality of the recording (Gannon et al. 2022). A solution could be to use an external microphone if the camera allows such input. While recording on camera, with or without an external microphone, the recordist should consider that: 1) microphone pre-amps in cameras typically have a much lower quality than external pre-amps; and 2) the camera's default recording parameters (e.g. gain) should be fine-tuned to the external microphone. Alternatively, when recording both video and sound separately, a sharp noise at the start of the recording will allow synchronization between the data streams coming from the camera and the sound recorder. While recording on camera or audio recorder, be it via microphone or hydrophone, it is always advisable to monitor the ongoing recording via headphones when possible.

9780192886491_Book.indb 544 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:01

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

29.2.2 Storage and metadata

Recordings can be stored in various formats; common file extensions include WAV (Waveform sound files), PCM (Pulse Code Modulation), and MP3. The last of these uses lossy data compression to reduce file size. In most cases, the operated compression only retains signals in a frequency band up to 16 kHz, therefore spectrum analysis using MP3 files is discouraged. On the contrary, uncompressed formats such as WAV or PCM are extremely reliable and, indeed, the WAV is undoubtedly the preferred extension. However, being uncompressed, WAV files can become heavy (one-hour recording sampled at 48 kHz and 24-bit requires about 0.5 GB of storage). Therefore, when recording for long periods, it is advisable to limit either the sampling frequency or resolution or, when possible, set the recorder to produce files of shorter duration (Gannon et al. 2022).

Associated data are as important as the recordings themselves. A good practice is annotating GPS coordinates, time of day, sex, age, and identity of the emitter, context of emission, and additional data that could be helpful in further analyses. In the past, manual annotation was the norm; today, devices can automatically track and store metadata. When recording in the wild, months could pass before being able to actually process the data. In either case, especially when recording duets or choruses, where at least two individuals are involved, we highly recommend to 1) store metadata and 2) pre-process the recorded audio files, for instance by using annotation tools (i.e. Praat TextGrid) directly in the field, as soon as possible after the data have been collected. Praat TextGrid files are particularly recommended because they allow batch processing and extraction of spectral and rhythmic features via Praat and Python (Boersma 2014; Jadoul et al. 2018, 2024; van der Werff et al. 2024).

It is advisable to store the data via hard or solid-state drivers (HD or SSD), internet clouds, internal servers (if possible). The same rules applied to the equipment apply here: drivers should be stored in dry and safe places. Guess what? If you record in a rainforest you will hardly have any internet access, let alone a dry place. What to do then? The only thing we feel confident to recommend is to create multiple back-up copies (what are the chances that three different drivers will be damaged?), store the back-up drivers with large quantities of desiccants (like silica gel), and upload the data via online repositories as soon as possible.

29.3 Analysis techniques

The plethora of analytical methods can be overwhelming. What analyses to use? Which features of signals should one focus on? For a long time, animal bioacoustics (and potentially also speech research) has focused on analysing features in the spectral dimension or element combinatoriality (Ravignani and Norton 2017). While these two are absolutely important, the temporal/rhythmic dimension has been neglected. In

9780192886491_Book.indb 545 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:05:01

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

humans and other animals temporal structure does encode copious information; in this analysis section, we will focus on methods for rhythm analysis (for a rhythm reporting framework, see Hersh et al. 2023).

Our question above therefore becomes: how to characterize, measure, and compare temporal structures? A first conceptual step is to distinguish between temporal structures within a pattern or between different patterns, where the choice of a particular analytical method may depend on the level and granularity of analysis (Figure 29.1 shows different rhythmic patterns in both cases).

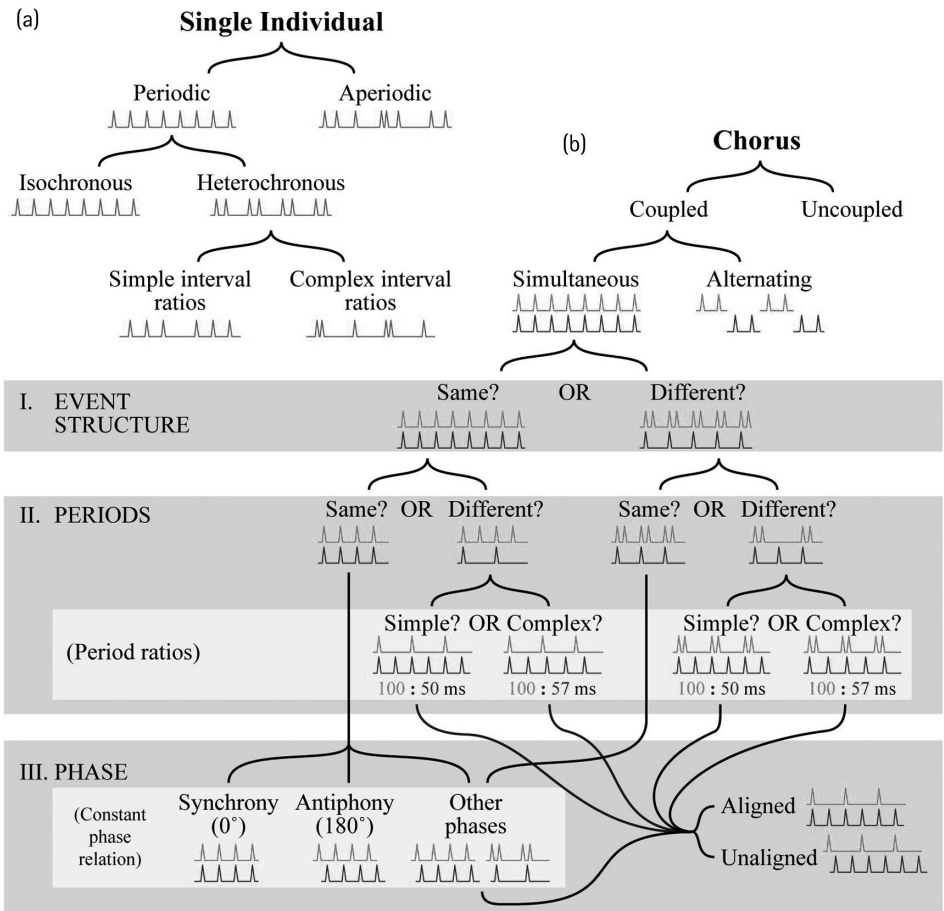


FIGURE 29.1 Potential, non-exhaustive representation of rhythmic patterns within a single individual and their potential relationship to other individual rhythmic patterns in a chorus.

For every abstract reproduction of a pattern, time flows from left to right and each peak represents and event in time. Different shades of grey represent different individuals.

Source: Ravignani et al. (2014). © Ravignani, Bowling, and Fitch 2014.
Reproduced under the terms of the Creative Commons Attribution License (CC BY).

When dealing with a pattern from a single individual, many techniques exist to quantify its internal rhythmic structure. A basic assessment of whether a signal is periodic

9780192886491_Book.indb 546 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:01

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

usually relies on a visual evaluation of the **inter-onset intervals** (**IOI**) between consecutive units within a sequence (e.g. Gamba et al. 2016). After visualization, quantifying the degree of periodicity is the next step. Burchardt and Knörnschild (2020) analysed short isolation monosyllabic calls of *Carollia perspicillata*, long isolation multisyllabic calls of *Saccopteryx bilineata*, and click trains of *Physeter macrocephalus*, and compared different analytical techniques. The authors found that the isochrony or heterochrony of a signal can be determined by measuring the **IOI** or the **normalized Pairwise Variability Index (nPVI)**, a measure of how variable consecutive intervals are (Figure 29.2), and that the **Kolmogorov–Smirnov** test can be used to investigate whether an interval distribution may differ from a given distribution (i.e. normal) if referring to an isochronous pattern (Ravignani and Norton 2017). Isochrony is the simplest form of periodicity and rhythmicity: the repetition of a signal with exactly the same IOI, like a metronome or the ticking of a clock. **Recurrence plots** can detect simple or complex heterochronicity. Exact beat frequencies can be computed via a **Fourier analysis (FFT)**, or a **generate-and-test approach (GAT)**, although the latter is interpretable only if sustained by a **goodness-of-fit value** (see also Burchardt et al. 2021 and Figure 29.2). Previous work has already reviewed strengths and weaknesses of the different techniques, so we will not discuss those here in detail (Ravignani and Norton 2017). The nPVI analysis condenses temporal regularities by providing a single number defining the variability of a series of IOI but is not robust across replications. The FFT is probably one of the most used and fastest methods that decompose the signal into isochronous pulses but may lead to multiple best-fitting values. A potential solution to this ambiguity is represented by the GAT, which finds the slowest pulse best fitting the pattern; however, this method is computationally demanding, therefore less common than the FFT.

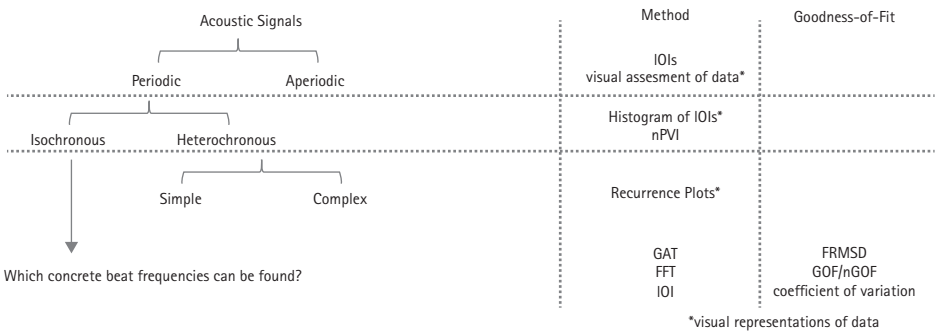


FIGURE 29.2 Methods (right side) used to test between possible features of a pattern (left side) and potentially place that rhythmic pattern in the taxonomy of Figure 29.1.

Source: Burchardt and Knörnschild (2020). © Buchardt and Knörnschild 2020. Reproduced under the terms of the Creative Commons Attribution License (CC BY).

Vocal rhythms can also be measured through **transition probabilities** between two given events in a sequence, i.e. the probability that one of the events is followed by the other (Ravignani et al. 2019a). **Autoregressive-moving average models** (**ARMA**) allow us to predict whether the value of a given IOI can be obtained via a linear combination of

9780192886491_Book.indb 547 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:05:02

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

previous interval values, but it only works with linear relations. **Recurrence plots** (Dale et al. 2011) and **chronotopological maps** (Jacoby and McDermott 2017) are extremely useful for visualizing temporal regularities. The former are suitable to visualize patterns between non-adjacent segments but are potentially prone to false positives or negatives; the latter visualize ratios between consecutive intervals in a bidimensional triangular simplex in which every three-interval rhythm is represented through a point on the simplex. Lastly, given that isochrony has been defined as a fundamental component of rhythmicity across cultures and species (Savage et al. 2015; Roeske et al. 2020), a recent branch of research is focusing on finding musical features consistently found across human cultures (the so-called '**musical universals**') across different non-human species as well. In this scenario, the use of **rhythmic categories** (Roeske et al. 2020; De Gregorio et al. 2021b; Demartsev et al. 2022; Raimondi et al. 2023; Ma et al. 2024; Nave et al. 2024) proved efficient in evaluating commonalities between humans and non-human animals.

The panorama of methods changes when examining more than one pattern (Figure 29.1). A recent paper showed how vocal interactive rhythmicity can be quantified using a multimethod approach (Anichini et al. 2023). A first, common method to analyse temporal relations across time series is the **Granger causality test**. Borrowed from econometrics (Granger 1969), it allows us to test whether the temporal intervals of a given series are useful in predicting those of a second series. The **ARMA** model represents an optimal solution for multidimensional time series. In this case, the model is used to predict whether the IOI of a given pattern can be measured through a linear combination of the intervals characterizing another sequence. Recently, the **Adaptation and Anticipation Model** (**ADAM**) has been employed to investigate temporal adaptation in vocal interactions in seals (Anichini et al. 2023). Further methods, such as the **circular statistics** and the **Allan Factor Analysis**, allow, respectively, verifying hypotheses on events periodically occurring, and measuring the grouping of events in time, through the quantification of their variances at different temporal scales (Kello et al. 2017).

29.4 Conclusions

To summarize, we have discussed different phases of bioacoustics research. We started with a brief overview of sub-areas of research and some key definitions and concepts. We discussed in detail how to perform sound recordings, with particular emphasis on the behaviour and ecology of a species: the bioacoustician working with multiple species needs to constantly adapt not only to the environment, but to the diverse behaviour of the organisms they are trying to record. We also discussed more technical aspects of bioacoustics, such as which microphone to choose or how to store data. We concluded with analysis techniques exclusively focused on the rhythmic and temporal dimension of animal sounds.

As the field progresses, we believe some sub-areas of bioacoustics will see particularly fruitful developments. Simultaneous recording of sound production, physiology, and

9780192886491_Book.indb 548 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:02

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

movement can deliver rich data streams, and their integration can serve to test a new generation of hypotheses, spanning from behavioural ecology to embodied cognition. Applying techniques from human voice science, phoniatrics, and logopedics to animals has delivered surprising results in terms of mechanisms: while some species use the exact same anatomical underpinnings we adopt for speech, others have evolved very different structures which can nonetheless function very much like the human larynx and upper vocal tract (Herbst et al. 2023; Madsen et al. 2023; Ravignani and Herbst 2023; Elemans et al. 2024). Non-invasive recordings of direct or indirect measurements of nervous system activities will provide a complementary mechanistic basis, potentially bridging a long-existing gap between biology, psychology, and neuroscience.

While performing or collaborating on a bioacoustics project, we believe it is important for all parties involved to understand every step of the process. Obviously, as in human speech and language, an animal sound unit does not exist in a vacuum. It is linked to the life history, behaviour, ecology, social network (to name just a few) of that particular individual. Because of this, we suggest that purely quantitative approaches should always be complemented by biological expertise. For instance, some apply advanced AI techniques (e.g. deep learning) to animal communication data while treating it like 'alien messages' of which we do not know anything else. While this can be potentially very fruitful, it needs to be complemented with knowledge on the behaviour and biology of that particular species.

In our experience, the field of bioacoustics is very open and welcoming to those who want to join. With this chapter, we hope to encourage new collaborations and synergies across disciplines.

Conflict of interest

The authors did not receive any benefit or incentive to support and recommend a particular brand.

References

Anichini, M., de Reus, K., Hersh, T. A., et al. (2023). 'Measuring rhythms of vocal interactions: A proof of principle in harbour seal pups', *Philosophical Transactions of the Royal Society B: Biological Sciences* 378(1875): 20210477, https://doi.org/10.1098/rstb.2021.0477.

Boersma, P. (2014). 'Acoustic analysis', *Research Methods in Linguistics* 375.

Browning, E., Gibb, R., Glover-Kapfer, P., and Jones, K. E. (2017). *Passive Acoustic Monitoring in Ecology and Conservation*. WWF-UK, https://doi.org/10.25607/OBP-876.

Burchardt, L. S. and Knörnschild, M. (2020). 'Comparison of methods for rhythm analysis of complex animals' acoustic signals', *PLoS Computational Biology* 16(4): e1007755, https://doi.org/10.1371/journal.pcbi.1007755.

9780192886491_Book.indb 549 /12_first_proofs/first_pr…/xml_for_typesetting 27-Mar-25 23:05:02

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Burchardt, L. S., Briefer, E. F., and Knörnschild, M. (2021). 'Novel ideas to further expand the applicability of rhythm analysis', *Ecology and Evolution* 11: 18229–18237, https://doi.org/10.1002/ece3.8417.

Clink, D. J. and Lau, A. R. (2020). 'Adherence to Menzerath's Law is the exception (not the rule) in three duetting primate species', *Royal Society Open Science* 7(11): 201557, https://doi.org/10.1098/rsos.201557.

Clink, D. J., Hamid Ahmad, A., and Klinck, H. (2020). 'Gibbons aren't singing in the rain: Presence and amount of rainfall influences ape calling behavior in Sabah, Malaysia', *Scientific Reports* 10(1): Article 1, https://doi.org/10.1038/s41598-020-57976-x.

Dale, R., Warlaumont, A. S., and Richardson, D. C. (2011). 'Nominal cross recurrence as a generalized lag sequential analysis for behavioral streams', *International Journal of Bifurcation and Chaos* 1(4): 1153–1161, https://www.worldscientific.com/doi/10.1142/S0218127411028970.

De Gregorio, C., Zanoli, A., Valente, D., et al. (2019). 'Female indris determine the rhythmic structure of the song and sustain a higher cost when the chorus size increases', *Current Zoology* 65(1): 89–97, https://doi.org/10.1093/cz/zoy058.

De Gregorio, C., Carugati, F., Estienne, V., et al. (2021a). 'Born to sing! Song development in a singing primate', *Current Zoology* 67(6): 597–608, https://doi.org/10.1093/cz/zoab018.

De Gregorio, C., Valente, D., Raimondi, T., et al. (2021b). 'Categorical rhythms in a singing primate', *Current Biology: CB* 31(20): R1379–R1380, https://doi.org/10.1016/j.cub.2021.09.032.

De Gregorio, C., Carugati, F., Valente, D., et al. (2022). 'Notes on a tree: Reframing the relevance of primate choruses, duets, and solo songs', *Ethology Ecology & Evolution* 34(3): 205–219, https://doi.org/10.1080/03949370.2021.2015451.

Demartsev, V., Haddas-Sasson, M., Ilany, A., et al. (2022). 'Male rock hyraxes that maintain an isochronous song rhythm achieve higher reproductive success', *The Journal of Animal Ecology* 00: 1–12, https://doi.org/10.1111/1365-2656.13801.

Dilungana, S., Deleforge, A., Foy, C., and Faisan, S. (2022). 'Geometry-informed estimation of surface absorption profiles from room impulse responses', *2022 30th European Signal Processing Conference (EUSIPCO)* 867–871, https://doi.org/10.23919/EUSIPCO55093.2022.9909667.

Duengen, D. and Ravignani, A. (2023). 'The paradox of learned song in a semi-solitary mammal', *Ethology*.

Ehnes, M. and Foote, J. R. (2015). 'Comparison of autonomous and manual recording methods for discrimination of individually distinctive Ovenbird songs', *Bioacoustics* 24(2): 111–121, https://doi.org/10.1080/09524622.2014.994228.

Elemans, C. P., Jiang, W., Jensen, M. H., et al. (2024). 'Evolutionary novelties underlie sound production in baleen whales', *Nature* 1–7.

Fitch, W. T. (2017). 'Empirical approaches to the study of language evolution', *Psychonomic Bulletin & Review* 24(1): 3–33, https://doi.org/10.3758/s13423-017-1236-5.

Gamba, M., Torti, V., Estienne, V., et al. (2016). 'The indris have got rhythm! Timing and pitch variation of a primate song examined between sexes and age classes', *Frontiers in Neuroscience* 10, https://www.frontiersin.org/articles/10.3389/fnins.2016.00249.

Gannon, W. L., Dunlop, R., Hawkins, A., and Thomas, J. A. (2022). 'Collecting, documenting, and archiving bioacoustical data and metadata', in C. Erbe and J. A. Thomas (eds.), *Exploring Animal Behavior Through Sound: Volume 1: Methods*, 87–110. Springer International Publishing, https://doi.org/10.1007/978-3-030-97540-1_3.

Granger, C. W. J. (1969). 'Investigating causal relations by econometric models and cross-spectral methods', *Econometrica* 37(3): 424–438, https://doi.org/10.2307/1912791.

9780192886491_Book.indb 550 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:02

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Gustison, M. L., Semple, S., Ferrer-i-Cancho, R., and Bergman, T. J. (2016). 'Gelada vocal sequences follow Menzerath's linguistic law', *Proceedings of the National Academy of Sciences* 113(19): E2750–E2758, https://doi.org/10.1073/pnas.1522072113.

Herbst, C. T., Prigge, T., Garcia, M., et al. (2023). 'Domestic cat larynges can produce purring frequencies without neural input', *Current Biology* 33(21): 4727–4732.

Hersh, T. A., Ravignani, A., and Burchardt, L. S. (2023). 'Robust rhythm reporting will advance ecological and evolutionary research', *Methods in Ecology and Evolution*.

Jacoby, N. and McDermott, J. H. (2017). 'Integer ratio priors on musical rhythm revealed cross-culturally by iterated reproduction', *Current Biology* 27(3): 359–370, https://doi.org/10.1016/j.cub.2016.12.031.

Jadoul, Y., Thompson, B., and De Boer, B. (2018). 'Introducing Parselmouth: A Python interface to Praat', *Journal of Phonetics* 71: 1–15.

Jadoul, Y., De Boer, B., and Ravignani, A. (2024). 'Parselmouth for bioacoustics: Automated acoustic analysis in Python', *Bioacoustics* 1–19.

Jarvis, E. D. (2019). 'Evolution of vocal learning and spoken language', *Science* 366(6461): 50–54, https://doi.org/10.1126/science.aax0287.

Kello, C. T., Della Bella, S., Médé, B., and Balasubramaniam, R. (2017). 'Hierarchical temporal structure in music, speech and animal vocalizations: Jazz is like a conversation, humpbacks sing like hermit thrushes', *Journal of The Royal Society Interface* 14(135): 20170231, https://doi.org/10.1098/rsif.2017.0231.

Knörnschild, M., Feifel, M., and Kalko, E. K. V. (2013). 'Mother–offspring recognition in the bat *Carollia perspicillata*' *Animal Behaviour* 86(5): 941–948, https://doi.org/10.1016/j.anbehav.2013.08.011.

Ma, H., Wang, Z., Han, P., et al. (2024). 'Small apes adjust rhythms to facilitate song coordination', *Current Biology*.

Madhusudhana, S., Pavan, G., Miller, L. A., et al. (2022). 'Choosing equipment for animal bioacoustic research', in C. Erbe and J. A. Thomas (eds.), *Exploring Animal Behavior Through Sound: Volume 1: Methods*, 37–85. Springer International Publishing, https://doi.org/10.1007/978-3-030-97540-1_2.

Madsen, P. T., Siebert, U., and Elemans, C. P. H. (2023). 'Toothed whales use distinct vocal registers for echolocation and communication', *Science* 379(6635): 928–933, https://doi.org/10.1126/science.adc9570.

Mathevon, N., Casey, C., Reichmuth, C., and Charrier, I. (2017). 'Northern elephant seals memorize the rhythm and timbre of their rivals' voices', *Current Biology* 27(15): 2352–2356.e2, https://doi.org/10.1016/j.cub.2017.06.035.

Melo, I., Llusia, D., Bastos, R. P., and Signorelli, L. (2021). 'Active or passive acoustic monitoring? Assessing methods to track anuran communities in tropical savanna wetlands', *Ecological Indicators* 132: 108305, https://doi.org/10.1016/j.ecolind.2021.108305.

Nave, K., Carrillo, C., Jacoby, N., et al. (2024). 'The development of rhythmic categories as revealed through an iterative production task', *Cognition* 242: 105634.

Petkov, C. I. and Jarvis, E. D. (2012). 'Birds, primates, and spoken language origins: Behavioral phenotypes and neurobiological substrates', *Frontiers in Evolutionary Neuroscience* 4: 12, https://doi.org/10.3389/fnevo.2012.00012.

Raimondi, T., Di Panfilo, G., Pasquali, M., et al. (2023). 'Isochrony and rhythmic interaction in ape duetting', *Proceedings of the Royal Society B: Biological Sciences* 290(1990): 20222244, https://doi.org/10.1098/rspb.2022.2244.

9780192886491_Book.indb 551 /12_first_proofs/first_pr/xml_for_typesetting 27-Mar-25 23:05:02

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Ravaglia, D., Ferrario, V., De Gregorio, C., et al. (2023). 'There you are! Automated detection of indris' songs on features extracted from passive acoustic recordings', *Animals* 13(2), https://doi.org/10.3390/ani13020241.

Ravignani, A. and Herbst, C. T. (2023). 'Voices in the ocean', *Science* 379(6635): 881–882, https://doi.org/10.1126/science.adg5256.

Ravignani, A. and Norton, P. (2017). 'Measuring rhythmic complexity: A primer to quantify and compare temporal structure in speech, movement, and animal vocalizations', *Journal of Language Evolution* 2(1): 4–19, https://doi.org/10.1093/jole/lzx002.

Ravignani, A., Bowling, D. L., and Fitch, W. T. (2014). 'Chorusing, synchrony, and the evolutionary functions of rhythm', *Frontiers in Psychology* 5: 1118, https://www.frontiersin.org/articles/10.3389/fpsyg.2014.01118.

Ravignani, A., Della Bella, S. Falk, S., et al. (2019a). 'Rhythm in speech and animal vocalizations: A cross-species perspective', *Annals of the New York Academy of Sciences* 1453: 79–98, https://doi.org/10.1111/nyas.14166.

Ravignani, A., Kello, C. T., de Reus, K., et al. (2019b). 'Ontogeny of vocal rhythms in harbor seal pups: An exploratory study', *Current Zoology* 65(1): 107–120, https://doi.org/10.1093/cz/zoy055.

Rhinehart, T. A., Chronister, L. M., Devlin, T., and Kitzes, J. (2020). 'Acoustic localization of terrestrial wildlife: Current practices and future opportunities', *Ecology and Evolution* 10: 6794–6818, https://doi.org/10.1002/ece3.6216.

Riedman, M. (1990). *The Pinnipeds: Seals, Sea Lions, and Walruses* (Vol. 12). Berkeley, CA: University of California Press.

Roeske, T. C., Tchernichovski, O., Poeppel, D., and Jacoby, N. (2020). 'Categorical rhythms are shared between songbirds and humans', *Current Biology: CB* 30(18): 3544–3555.e6, https://doi.org/10.1016/j.cub.2020.06.072.

Savage, P. E., Brown, S., Sakai, E., and Currie, T. E. (2015). 'Statistical universals reveal the structures and functions of human music', *Proceedings of the National Academy of Sciences* 112(29): 8987–8992, https://doi.org/10.1073/pnas.1414495112.

Shannon, C. E. (1949). *Communication in the Presence of Noise | IEEE Journals & Magazine | IEEEXplore,* https://ieeexplore.ieee.org/document/1697831.

Tønnesen, P., Gero, S., Ladegaard, M., et al. (2018). 'First-year sperm whale calves echolocate and perform long, deep dives', *Behavioral Ecology and Sociobiology* 72(10): 165, https://doi.org/10.1007/s00265-018-2570-y.

Torres Borda, L., Jadoul, Y., Rasilo, H., et al. (2021). 'Vocal plasticity in harbour seal pups', *Philosophical Transactions of the Royal Society B: Biological Sciences* 376(1840): 20200456, https://doi.org/10.1098/rstb.2020.0456.

Torti, V., Gamba, M., Rabemananjara, Z. H., and Giacoma, C. (2013). 'The songs of the indris (Mammalia: Primates: Indridae): Contextual variation in the long-distance calls of a lemur', *Italian Journal of Zoology* 80(4): 596–607, https://doi.org/10.1080/11250003.2013.845261.

Valente, D., De Gregorio, C., Favaro, L., et al. (2021). 'Linguistic laws of brevity: Conformity in *Indri indri*', *Animal Cognition* 24(4): 897–906, https://doi.org/10.1007/s10071-021-01495-3.

van der Werff, J., Ravignani, A., and Jadoul, Y. (2024). 'thebeat: A Python package for working with rhythms and other temporal sequences', *Behavior Research Methods* 1–12.

9780192886491_Book.indb 552 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:02

CHAPTER 30

HOMININ TOOL EVOLUTION AND ITS (SURPRISING) RELATION TO LANGUAGE ORIGINS

RONALD J. PLANER, ELISA BANDINI, AND CLAUDIO TENNIE

30.1 Introduction: Varieties of cultural learning

It is now widely agreed that humans' cultural capacities were a crucial evolutionary driver of our species' unique success. However, it is also agreed that many other species have culture (e.g. van Schaik 2010). Yet, the key differences between the cultural capacities of humans and other animals remains a hotly disputed territory. Rather than focus on the controversies, our goal in this chapter is to set out a novel social learning framework (Bandini and Tennie 2018; Bandini et al. 2020), along with a working hypothesis within this framework. We then use this framework and hypothesis to explore a major unresolved issue in the evolution of language, namely, the modality and type of the very earliest forms of language—we think, with some surprises.

The signature feature of the framework we adopt concerns how it distinguishes various types of cultures. Culture *simpliciter* is conceived of liberally: whenever social learning of any kind plays any role in the distribution of a trait in a population, we have a case of culture—'minimal culture', *at least*[1] (Neadle et al. 2017). Social learning is then always cultural learning. This liberal conception is useful as it ensures that no

[1] This can include negative traits, e.g. *not* eating a certain food (know-what-not).

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

potentially enlightening cases of animal behaviour or artefacts are ruled out simply by definition. In our view, the more comparative evidence we have to go on, the better—and that includes (perhaps especially) marginal cases of cumulative culture (Buskell and Tennie 2021). In contrast to earlier frameworks, which distinguish types of cultural learning based on abstract, historically grown social learning mechanisms,[2] the present framework focuses on the type of information/knowledge that is socially transmitted. These types correspond to (e.g. English-language) question words, such as 'who', 'what', 'where', 'when', and 'how'. This conveniently allows us to talk of 'know-when cultures' (e.g. a know-when-to-hide culture), 'know-what cultures' (e.g. a know-what-to-eat culture), and so on.

Such a content-centred framework encourages us to search for the specific type(s) of socially transmitted information that underlies a given case of culture. In particular, it alerts us to the fact that socially learned information might result in an outward behaviour or artefact which is itself not socially learned. To give a very simple example: a chicken might socially learn *where* to peck, but the action of pecking itself is not socially learned (chicks peck). Implicit in this example is another crucial point: just because one type of information has been socially transmitted (e.g. know-where-to-peck), that does not necessitate that other types of information are socially transmitted (e.g. know-how-to-peck), although it may first *appear* that way (Tennie 2023).

This framework is useful for formulating hypotheses about what makes human cultural capacities special among apes. Some species may be proficient or even prodigious social learners in some of these content domains, but not others. Here, we take as a working hypothesis the idea that humans possess unique abilities for know-how social learning, at least relative to our great ape cousins (Tomasello 1999). By 'know-how', we mean information that constitutes certain behaviour and/or artefact forms. We focus here especially on tools, which we conceive of as material objects that agents use in goal-directed ways in interaction with their environments (Kelly 2013). The form a tool takes can be dependent on know-how social learning, alongside the specific ways a tool is manufactured, maintained, or used.

It is clear that chimpanzees, at least, can be trained to socially learn novel forms of know-how (e.g. Pope et al. 2018). The question is whether/to what extent wild, unenculturated great apes socially learn such information (Tennie 2019; Tennie et al. 2020). Putative cases of know-how social transmission among wild, and captive unenculturated chimpanzees must be examined carefully. In our view, at least some, if not many, of these cases are more parsimoniously explained as instances in which *other* types of social information are being transmitted (as in our toy example above). Yet other cases are more complicated, and some claim that these show a necessity for cultural transmission of know-how. Types of termite fishing and nut-cracking in chimpanzees are such debated cases (Boesch et al. 1994; Sanz et al. 2009). These debates continue (see

[2] And still poorly understood from a cognitive neuroscience standpoint, it should be added (Heyes 2018).

9780192886491_Book.indb 554 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:02

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

e.g. Koops et al. 2022; Tennie and Call 2023)—though we note that the number of even candidate cases seems to shrink over time (Motes-Rodrigo and Tennie 2021).

The question for a necessity of know-how transmission brings up another issue. We need to further divide the category of know-how social learning. One intuitive distinction concerns the nature of the know-how that is socially transmitted. Is this know-how something that a species-typical, suitable individual might have learned on their own (with a non-trivial probability), given the appropriate conditions, e.g. environmental scaffolding/developmental stages and/or motivation levels (see also Bandini et al. 2020)? If so, such know-how can be said to lie within the species' *zone of latent solutions* (ZLS) (Tennie et al. 2009, 2020). If not, it can be said to be an instance of *supra-individual know-how* (e.g. Tennie 2023).

We hypothesize that humans show a strong propensity for the social transmission of supra-individual know-how, relative to other great apes. This idea can be formulated in a stronger or weaker way. The strongest version would hold that *only* human apes are capable of such social transmission. A weaker version might hold that, while unenculturated, untrained apes can perhaps engage in supra-individual know-how social learning, there are important constraints (frequency/depth) on such learning in them, whereas the human capacity for such learning is frequent and wide. Here, we work with the strongest version, as we think (i) it makes for a clearer exposition, and, importantly, (ii) it is consistent with current evidence. That said, should the evidence ultimately refute this version, the weaker versions of the hypothesis would remain viable (see also Reindl et al. 2018).

Terminology is a necessary evil. Note that the terms 'know-how triggering' (of latent solutions) and 'know-how copying' (of supra-individual know-how) have been used before to describe different types of know-how cultural learning as related to the ZLS (Tennie 2023). We shall follow that convention here, as they abbreviate these relationships. However, note that these terms abstract away from the underlying neurocognitive details of learning.[3] If, as we hypothesize, there is a real difference between humans and (say) chimpanzees[4] in the ability to socially learn supra-individual know-how, this must be in virtue of underlying neurocognitive differences between these species. It cannot simply be a brute fact.[5]

We do not mean to imply that our abilities for know-how *copying* constitute the total story as to what makes human cultural learning unique. In particular, even if other great apes are proficient social learners in each of the other content domains, the *bandwidth* of human cultural learning in these domains seems considerably larger. As groups, we are capable of accumulating and transmitting vastly more information across these

[3] We take this term to encompass motivational features.

[4] Unless otherwise specified, here and in the following we will always mean unenculturated, untrained primates.

[5] Very likely, these differences concern the capacity and motivation to parse behaviour/artefact forms in fine-grained ways, and to match the components and organization of these forms. Note that these differences can be realized ontogenetically.

9780192886491_Book.indb 555 /12_first_proofs/first_pr…/xml_for_typesetting 27-Mar-25 23:05:02

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

domains. This very likely reflects not just differences in our intrinsic cognitive make-ups, but also differences in how we engineer our learning environments (i.e. niche construction) (Sterelny 2012; Laland et al. 2016). Moreover, there are co-evolutionary interactions among these domains: as one's know-how repertoire expands—as one comes to be able to *do* more—more information becomes relevant and worth socially learning. Also, more supra-individual know-how can be generated. Culture breeds culture.

All that said, our abilities for know-how copying do play a special role in explaining how it is that we can culturally evolve behavioural and artefact forms of extreme complexity and/or arbitrariness. Rituals, dances, and art forms are all paradigm examples. And so, too, are human languages, which make up our other major focal point in this chapter. There are numerous respects in which human languages go well beyond our species' ZLS. Human language words are the most obvious way in which this is true, and probably also the most striking (Motes-Rodrigo and Tennie 2021). Each word is a behavioural form unto itself, and the average speaker knows tens of thousands of words. Virtually all spoken words are arbitrary, while even sign languages, which make comparatively heavy use of iconicity, still contain many arbitrary signs. There is simply no way for an individual to learn on their own the know-how corresponding to such words. Rather, it must be directly copied from other language users.

We shall return to connections between know-how copying and language in due course. First, we discuss some methods that help us to identify know-how copying. Then, we move on to the hominin technological record, with an eye towards (very roughly) dating the origins of know-how copying in our line. The timeline we are led to infer places a substantive and unobvious constraint on the origins of language in our line. Taking this constraint seriously, in our view, pushes one in the direction of a gestural–iconic origins scenario for language evolution, effectively providing an additional line of support for a gestural–iconic origins theory over its vocal–arbitrary origins competitor. In our view, this provides a proof of concept of how the methods for identifying know-how copying which are outlined in this chapter can be leveraged for understanding language evolution.

30.2 Identifying know-how copying

What tools do we have for recognizing know-how copying? Let's first take a step back and ask how we can identify a culture (including minimal kinds) at all. One common method for doing so is the *method of exclusion* (Whiten et al. 1999). This method involves looking for differences in trait distributions across populations and attempting to exclude genetic and environmental difference explanations. If these alternatives are deemed to be ruled out, then social learning is adopted as the explanation behind the behaviour. This method can be useful for identifying likely cultures, but alone, it cannot tell us what type of information is socially transmitted. It lumps the many types of information together (e.g. Acerbi et al. 2022).

9780192886491_Book.indb 556 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:02

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

So, to home in on know-how copying, other methods are needed. One method is to run so-called *cultural island tests* (Tomasello 1999; Tennie et al. 2016). These studies involve providing suitable (e.g. motivated) naive individuals with all the material and context of a behaviour, but no information on the know-how. Both non-human primates and modern humans can serve as interesting and appropriate participants in experiments of this kind (though, obviously, the significance of how they perform may differ depending on their species and individual participants' backgrounds). If re-innovation of the know-how can be elicited, this provides proof that it lies within the species' ZLS, and so we can conclude that know-how copying is not at work (although other types of social learning may be).

Another method is to use direct *know-how copying tests* (Tennie et al. 2009). The logic behind such tests is straightfoward. If it is indeed the case that, say, chimpanzees, engage in know-how copying (e.g. in the wild), then they ought to be able to socially learn at least some other forms of supra-individual know-how under ecologically valid learning conditions.[6] So, what we do here is model such a form that proved supra-individual in baseline (e.g. a looped piece of string as a tool) for a group of individuals and see whether they in fact succeed in reproducing it.[7] To the extent that individuals show repeated copying failures (especially over a wide range of forms), then this is good evidence that they lack the ability to know-how copy spontaneously (without human intervention). Indeed, apes fail this test (review in Tennie et al. 2020).

Still another method is known as the *method of local restriction* (Motes-Rodrigo and Tennie 2021) (Box 30.1). The main idea here is that know-how copying should eventually give rise to behaviour/artefact forms that are unique to a single cultural population. Other things being equal, this methodological approach assumes a population of learners who are capable of know-how copying to quite rapidly evolve forms that escape their species' ZLS by way of cumulative additions or refinements of extant forms. Except, perhaps, during the earliest stages of such a process, or when task demands essentially require a certain design, it is unlikely that the same supra-individual know-how form should evolve in two informationally isolated populations. So, restriction of a behavioural/artefact form in space and time to a single population or lineage is good (though still defeasible) evidence for know-how copying.[8]

By using a combination of these methods, we can develop a clearer picture of who has know-how copying, and who does not. In our view, the current body of evidence (which is vast) suggests that, among great apes, only humans possess this ability. Known great-ape behavioural/artefact forms have been shown to repeat

[6] In particular, highly artificial forms of reinforcement should not be necessary.

[7] Obviously, the form chosen must be sensible. It should not be something that is difficult for the relevant individuals to learn due to physical constraints. It can also consist of behavioural forms (Tomasello et al. 1997; Tennie et al. 2012; Clay and Tennie 2018; Neadle et al. in prep).

[8] Still defeasible because the form might still ultimately be explained in terms of the social transmission of another information type or know-how triggering. That possibility becomes less likely, however, the more complex/arbitrary the form is.

9780192886491_Book.indb 557 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:05:02

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Box 30.1 The method of exclusion

A classic method to study culture (the 'method of exclusion') seeks to identify cases in which the presence of a behaviour (or artefact) in a population cannot be explained via genetic or environmental factors. These cases, by method of exclusion, are then deemed to be cultural. If, for example, a tendency to eat food A over food B is found in one population, but a tendency to eat food B over food A is found in another population, and no obvious genetic or environmental factors can account for these tendencies, then it is reasonable to conclude that these patterns are cultural (or more precisely, they would be patterns of a food-choice culture). Whilst a general critique of this classic method exists (e.g. Laland and Janik 2006), a more specialized critique has often been raised against it, namely that this method cannot pinpoint which type(s) of social learning underlie the patterns it detects (summary in Acerbi et al. 2022).

The method of local restriction

The method of local restriction aims to identify these underlying social learning mechanisms. This method was named and formalized by Motes-Rodrigo and Tennie (2021), though its underlying logic is based on previous work (see e.g. Byrne 2007; Whiten et al. 1999, 2001). The foundational premise of this method is that iterated pathways of know-how copying should be expected to eventually give rise to *population-specific* types of know-how (e.g. due to accumulated copying errors). True cases of population-specific know-how of this kind can then be used as (indirect) evidence for specific types of social learning (i.e. know-how copying). Consequently, the method of local restriction is designed to look for potential cases of population-specific types of know-how (which must then, however, pass various additional criteria before they are categorized as such).

As we discuss below, the method of local restriction is particularly useful for identifying know-how copying in the archaeological record.

across isolated populations and would appear to be capable of re-innovation by naive members of the relevant species (Bandini and Tennie 2017; Motes-Rodrigo and Tennie 2021). Additionally, great apes fail direct tests of know-how copying (Tennie et al. 2012, 2020; Clay and Tennie 2018; Neadle et al. 2021). This is not to say that other great apes (or at least *Pan*) are incapable of other types of social learning; they clearly culturally learn in other ways. It appears though that know-how copying eludes them.

Put otherwise: the above methods enable us to form a highly empirically grounded and testable account of the social learning abilities of our earliest hominin ancestors. This account can and should influence how we read the archaeological record. The evidence suggests that know-how copying was not yet in place at the time of divergence with *Pan*. However, we know that at some point, our ancestors went on to evolve extremely powerful know-how copying skills. When might that have happened? And how can we tell?

9780192886491_Book.indb 558 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:02

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

30.3 Know-how copying and the archaeological record

How can we identify know-how copying in the hominin past? Direct copying and cultural island tests require active participants. However, in the case of the latter, other great apes or modern humans may sometimes serve as appropriate stand-ins for ancient hominins (Rolian and Carvalho 2017; Bandini et al. 2022). If, for example, chimps can re-innovate a certain archaeologically known tool form, then that is rather good evidence that this form was in the ZLS of its hominin maker. This has in fact been shown in principle for the very earliest flaked tools in orangs (Motes-Rodrigo et al. 2022). Similarly, if a modern human who is completely ignorant of stone tools can re-innovate a certain tool form, then this provides additional evidence for thinking that the form was in the ZLS of its hominin maker. Experiments like this have been done, too—with positive results (Snyder et al. 2022). Obviously, any hominin conclusions reached on the basis of human tests will be criticized. It is very hard to rule out completely the possibility that participants benefited from previous knowledge (though to the best of our ability this was ruled out in the study). Additionally, their brain sizes and ontogenies are fully modern.[9] Still, these results are highly intriguing, and they question the (quasi) status quo opinion of seeing even the earliest stone tools in the record as being based on know-how copying. To make a strong case, additional data would be beneficial

Fortunately, the method of local restriction is quite well suited to help in the task of identifying likelihoods of know-how copying in the archaeological record. In what follows, we give a brief overview of how the evidence looks from the perspective of this approach.[10]

From about 150 thousand years ago (kya), we begin to see multiple 'smoking gun' traces for know-how copying in the archaeological record. These traces include the production of material symbols, such as forms of jewellery and engraved designs, that *do* seem restricted to certain cultural populations (Henshilwood et al. 2002; Conard 2005; Lombard 2012; Texier et al. 2013; Wadley 2015).[11] They also include novel tool forms that are more obviously cumulative[12] in character *per se*: for example, projectile weapons, such as the atlatl and bow-and-arrow; fire-bows; microliths and bladelets, etc. (Shea 2009; Sahle and Brooks 2019; Lombard and Shea 2021). The evidence is *strong* that, by at least 80 kya, our ancestors were leading cultural lives on par with those of ethnographically known foragers.[13] They were *behaviorally modern* peoples.[14]

[9] The use of very young children as participants (as suggested by the authors) can damp down these problems, but only partially.

[10] For a fuller treatment, see Tennie (2023).

[11] Or at least *meta*-groups. Here we have in mind groups that interact, and hence exchange information, with each other.

[12] But also more obviously supra-individual.

[13] Who *all* have know-how copying skills (for example, evidenced by *ubiquitous* language use).

[14] Another compelling argument in support of this timeline is that it is around this time that modern humans began to disperse from Africa throughout the rest of the world. Behavioural modernity did not evolve independently in all these different groups.

9780192886491_Book.indb 559 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:05:02

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Thus, there is no question that our ancestors had evolved modern skills for know-how copying by *at least* the early Late Pleistocene or late Middle Pleistocene. How much earlier might these skills have evolved?

Any artefacts our ancestors made out of soft/organic materials are unlikely to survive for periods of hundreds of thousands of years, never mind millions of years. The earliest such artefacts we know of are wooden spears found at Schoningen, Germany (Thieme 1997). These date to around 400 kya, and the preservation conditions at this site are unusually good. Given that chimpanzees make wooden tools, we can be relatively sure that our earliest ancestors did so as well (actually already before the split with the last common ancestor). However, it is almost certain that such soft-material technologies are forever lost to us.

Therefore, we need to rely almost exclusively on stone-tool evidence as a guide to potential know-how copying in the deeper hominin past. The earliest suspected stone tools come from East Africa and are claimed to date to around 3.3 mya (Harmand et al. 2015). They are extremely primitive flaked tools, and it is unlikely that their makers had in mind a particular design (above producing a useable cutting edge). In addition, there are claims for cut-marked bones at another East African site around the same time (McPherron et al. 2010). Whether or not these dates are correct, things largely remain this way for the next million years or more: flakes with cutting edges. The best known grouping is the so-called *Oldowan* (though some archaeologists make sub-classifications).[15]

In our view, the tools from this period do not require an explanation in terms of know-how copying. They were (almost certainly) in the ZLS of their hominin makers (australopithecines and *Homo habilis*). Evidence from cultural island tests performed with orangutans and modern humans supports this conclusion (Motes-Rodrigo et al. 2022; Snyder et al. 2022). This idea is further backed by evidence of simple flake production (Westergaard and Suomi 1994; Proffitt et al. 2016) and use (Westergaard and Suomi 1994) in capuchin monkeys. In addition, these tool forms appear in different populations throughout Africa for which, minimally, we lack evidence of social connectedness (Davidson 2016). Finally, and perhaps most importantly, we think there is no compelling evidence of directional or cumulative change. This is not to say there is no variation whatsoever over space or time in Oldowan assemblages. Rather, it is to say that what variation *does* exist is (arguably) fully explainable in terms of differences in raw material types, species (e.g. anatomies) and noise inherent to manufacture

[15] In particular, some theorists recognize both an *Oldowan proper* and *Developed Oldowan* groupings. On the other hand, other theorists simply regard the Developed Oldowan as part of the early Acheulean. For an (empirical) discussion of such issues, see e.g. Semaw et al. (2009) and Stout et al. (2010). These finer ways of carving up the Oldowan do not matter for the argument developed here. That is true even if later Oldowan tools made somewhat increased cognitive demands (e.g. on mental rotation) as compared to the earliest tools from this industry. Hence, we set these finer distinctions aside in what follows.

9780192886491_Book.indb 560 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:02

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

processes.[16] There is no signal of incremental change that would suggest copying know-how chains.

Then, around 1.8 mya, a new form of technology appears. This is the beginning of the so-called *Acheulean industry*. This industry is often epitomized by its large, oval shaped, and bifacially flaked hand-axes. The invention of these tools (and other large-cutting tools, e.g. cleavers) have been plausibly interpreted as signalling a shift to a lifeway in which meat, and butchering, was more central (Shea 2017; Toth and Schick 2019). Unlike Oldowan tools before them, most experts think that their makers (*Homo erectus*, and later, *Homo heidelbergensis*) *did* have a particular form in mind during manufacture—that they worked from a mental template of some kind.

Acheulean hand-axes are no doubt impressive, and in contrast to Oldowan tools, they are challenging—even for modern humans—to make (though the early ones are perhaps not very challenging at all; compare Moore and Perston 2016). Yet, *Homo erectus* managed to reliably reproduce this form over long stretches of time and space.

Surprisingly, however, they are not clear indicators that know-how copying had already evolved in our lineage in any way. First, hand-axes appear in multiple parts of the world in populations seemingly culturally isolated from one another. In particular, hand-axes are known from parts of China where the next closest hand-axe producing population lived some 1,500 km away (Zhang et al. 2010; Corbey et al. 2016). Such re-innovation (if it is re-innovation) suggests that hand-axes were in fact in *Homo erectus*' ZLS. Setting this aside, the extreme conservatism of the Acheulean over vast stretches of space and time plausibly argues *against* transmission via know-how copying, rather than in favour of it, as many theorists have either explicitly or implicitly supposed (Foley 1987; Richerson and Boyd 2008; Tennie 2023). That is not the pattern we would expect to find if the technology was, in fact, culturally sustained via know-how copying. Recall from above that, other things being equal, we expect know-how copying to produce cumulative changes to behavioural/artefact forms, in turn producing forms whose complexity or arbitrariness (or both) exceeds what any individual could realistically invent on his or her own. Corollarily, where two or more populations of cultural learners are each engaged in repeated know-how copying, and do not exchange information with each other, we expect to see a clear divergence in forms.

Perhaps this is a case where other things are *not* equal. It has been suggested, for example, that *Homo erectus* possessed suprahuman conformist social learning; compared to us, they were supposedly *much more* concerned with matching the products of their peers (Finkel and Barkai 2018). Alternatively, it has been suggested that efficiency considerations relating to manual handling, raw materials, and transport strongly select for the conserved hand-axe design (see e.g. Ludwig and Harris 1998; Key and Lycett 2019; Fedato et al. 2020). To borrow some terminology from Dan Dennett (1995): on this sort of view, hand-axe features are, if not 'forced moves', then certainly 'good tricks' in design space. We are sceptical, though a lack of space prevents us from explaining our

[16] The term 'variable sameness' (Isaac and Isaac 1977) is sometimes used for such variation.

9780192886491_Book.indb 561 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:05:02

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

reasons for scepticism here.[17] Minimally, a lack of know-how copying provides a much more parsimonious explanation for Acheulean stasis.

By around 1 mya to 800 kya, there are hints that changes in know-how social learning were perhaps afoot (Kuhn 2020). This period is sometimes called the *developed Acheulean*. Hand-axes (and other large cutting tools) become more symmetrical and show extensive bifacial thinning. These changes in artefact form imply changes in manufacture technique (including, *likely*, novel meta-tools). In particular, they imply increased 'platform preparation', in which certain flakes are removed from a piece of worked stone *in order to set up* the removal of other flakes.[18]

In the ensuing several hundred thousand years, platform preparatory techniques become more varied, eventually giving rise to the so-called *Levallois industry*. Levallois tools typically consist of such lengthy sequences of preparatory flake removals that the finished artefact form could be released from the worked core with a single, final stroke of a hammer. It is also around this time that we begin to see the first evidence of composite tool technologies (Wilkins and Schoville 2012; Wilkins and Chazan 2012). Hafts (i.e. handles) themselves do not show up in the archaeological record until much later, but hafting can be inferred based on the (small) size of lithic blades and/or use and wear marks indicative of hafting.

There is much more to be said on these topics, but the above will have to suffice for now. Here is an outline of the picture that emerges. By around the beginning of the late Pleistocene, evidence for know-how copying in our line is overwhelming. This gives us a (hard) lower bound on the evolution of this capacity. As for an upper bound: Oldowan technology is most plausibly understood as lying within the ZLS of its hominin makers. More controversially, we think the same is, or at least very well might be, true of the early and middle portions of the Acheulean.[19] As we look forward to the late or developed Acheulean, and more thoroughly, Middle Pleistocene technologies, such as Levallois tools and composite tools, we enter an epistemological grey zone. Here, claims for the evolution of know-how copying strike us as plausible, but are much harder to demonstrate than many theorists often realize. More work on the connection between these technologies and the evolution of know-how copying is needed.

For the purposes of this chapter, it is really only the first third of this timeline that matters, however. In particular, if *Homo erectus* living before, say, 1 mya indeed lacked skills for know-how copying, then this provides a surprising and interesting constraint on theories of language evolution.

[17] For a more detailed critical response to these alternative proposals, see Tennie (2023). This article also responds to views that reject the claim of conservatism here.

[18] As opposed to each flake removal directly contributing to the final artefact form. That process is known as *marginal trimming*, and results in much 'chunkier' tools. Note that elaborate platform preparation implies higher general cognitive skills (e.g. enhanced forms of planning)—but does not in itself pinpoint copying.

[19] So, from about 1.8 to around 1 mya.

9780192886491_Book.indb 562 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:02

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

30.4 Lessons for language evolution: The curious case of *Homo erectus*

Let us summarize the discussion so far. We began by distinguishing various types of cultural learning based on the type of information that is transmitted: know-what, know-where, etc., and know-how. We then described some methods for identifying genuine cases of know-how cultural learning (as opposed to other types). The task is not always an easy one, but it is empirically tractable. Next, we saw how evidence produced by these methods (and in some cases, just the general logic behind them) can help us interpret the archaeological record. In particular, they allow us to develop an empirically grounded hypothesis as to when know-how copying established in the line leading to humans. As we now discuss, the timeline suggested by this evidence interacts with theories of language origins in some interesting ways.

The action centres on *Homo erectus*. This species evolved in Africa around 1.8 mya, is widely seen as much more human than any hominin species that came before. *Homo erectus* brains are twice the absolute size of present-day *Pan* brains. At least some of their life history patterns had shifted in the direction of modern humans; in particular, their rate of physical maturation slowed relative to their rate of brain development (Maslin 2016). A prolonged juvenile stage has been plausibly seen as an adaptation to a lifeway more demanding of social learning (Richerson and Boyd 2020). But delaying physical maturation in this way—and hence time to reproductive maturity—has both costs and risks. *Homo erectus* young would have remained dependent for longer. And they would have incurred an increased risk of dying before reproducing. Very plausibly, to meet these challenges, *Homo erectus* practised some form of reproductive cooperation, i.e. biparental care and/or cooperative breeding (Hrdy 2009; Opie and Power 2009; Van Schaik and Burkart 2010). A cooperative breeding regime can be expected to reduce extrinsic mortality risks, which, in turn, reduces the risks of growing up more slowly.[20] Most obviously, dependent young are less likely to starve, and are probably also at a reduced risk of predation and other harms (Planer 2023). At the same time, increased allomaternal care likely enabled *Homo erectus* mothers to begin to reduce their interbirth intervals. By having more offspring, a mother can 'hedge the bet' she makes in having offspring that mature more slowly. One of the most striking features of *sapiens* lifeways is that we develop more slowly than any other great ape while spacing our births more closely together than any other great ape (Robson and Wood 2008; Pilbeam and Lieberman 2017; Nakahashi et al. 2018). This trend likely began in *Homo erectus*. In addition, reproductive cooperation can be expected to forge enduring social bonds which, when combined with dispersal of individuals from their natal group to nearby residential groups, facilitates intergroup tolerance, and hence

[20] See Gutherie (2007) in the connection between such risks and life history evolution.

9780192886491_Book.indb 563 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:05:02

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

intergroup information exchange (Chapais 2008; Planer 2023). Each of these factors, and especially their sum, probably played an important role in enabling *Homo erectus* to leave Africa and rather rapidly colonize much of the Old World, hanging on in some places until strikingly late dates.[21]

There are also converging lines of evidence that *Homo erectus* showed new types of collaborative foraging (Planer and Sterelny 2021). Above, we mentioned *Homo erectus*' larger and hence more metabolically expensive brains. *Homo erectus* bodies also changed in ways that are indicative of a significant dietary shift. While they generally grew larger overall, *Homo erectus* had smaller and less bell-shaped ribcages, and also smaller intestinal tracts ('guts') relative to earlier hominins. They also show a reduction in craniofacial and dental anatomy (Wrangham 2017). These changes have been plausibly seen as reflecting an increased role for animal products (meat, marrow, brains) in *Homo erectus* diets.

But how, exactly, did *Homo erectus* secure greater quantities of such products for themselves?[22] Some were very likely scavenged from carcasses which predators, such as big Pleistocene cats, had abandoned after having had their feed. Such meat could have laid exposed to the elements—in particular, the sun—for a prolonged period of time, and thus may have begun to putrefy. It is tempting to think (and many have thought) that our ancestors would have tended to avoid eating putrefied meat due to the risks of contracting food-borne illnesses, but there is ethnographic evidence that challenges this idea (see e.g. Speth 2017; Speth and Morin 2022). If—and to the extent that—putrid meat can be consumed at tolerable risk levels, there is even *benefit*: putrefaction breaks down the animals, making them easier to digest, and hence of higher net energy yield. In addition, *Homo erectus* would have consumed a wide range of animals, some opportunistically, some targeted. This would have included insects[23] (Lesnik 2018), as happens in the vast majority of ethnographically known foragers (Kelly 2013), and in many contemporary primates (Rothman et al. 2014).[24]

[21] The last known *Homo erectus* specimen comes from Java and has been dated to between 117 and 108 kya (Rizal et al. 2020).

[22] We here leave aside (due to a lack of space) the important question of *why Homo erectus* sought to incorporate more animal products into their diets. It is standard to think in terms of caloric payoffs when attempting to understand such decisions. However, there is more to the dietary equation than caloric intake. In particular, primates need to satisfy certain micronutrient (i.e. vitamin, mineral) needs for good overall health. Animal products, even in very small quantities, can provide them with such crucial micronutrients, and may be targeted, even when individuals might obtain a higher caloric payoff by other means. Tennie et al. (2009) call this idea the 'meat-scrap hypothesis', which they offer as an explanation for the group hunting behaviours of chimpanzees (and possibly other social omnivores). On this hypothesis, it can make sense for a chimpanzee to join in a group hunt, even if he can at best expect to receive only a very small portion (a scrap) of meat from a hunt. A similar role for animal products in the diet of early hominins was proposed by Milton (1999).

[23] The maggots in putrid meat included, if *Homo erectus* in fact ate putrid meat to the point of maggot stages.

[24] While both vertebrate and invertebrate resources were likely important, the micronutrient (vitamins, minerals) profile of these two resources can differ. Vertebrates may be preferentially targeted

9780192886491_Book.indb 564 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:02

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

But the archaeological evidence suggests that *Homo erectus* were more ambitious in their animal-getting efforts. Specifically, between 1.8 and 1.6 mya, we begin to find bone assemblages from medium and large animals in association with Acheulean tools in East Africa. Many of the bones show defleshing scars, indicating not only tool use, but that meat in need of butchering still remained on the bones at the time of carcass access. Very likely, *Homo erectus* engaged in so-called *power* or *bully scavenging*, actively driving away predators from kill sites, perhaps through some combination of hurled objects, targeted movements, and body positions and loud vocalizations. However, a number of theorists think that these assemblages actually evidence *even more*: big game hunting among *Homo erectus* (see, in particular, Pickering and Bunn 2012; Dominguez-Rodrigo and Pickering 2017). For many of the bones in these assemblages come from adult males in their prime (which have larger fat deposits). This is not the pattern we would expect to see if *Homo erectus* were limited to pirating carcasses from other top-guild predators; instead, we would then expect for the bones to predominately come from young or old animals (i.e. the easy-to-catch).[25] In what follows, we *assume* that (early Pleistocene) *Homo erectus* did in fact hunt medium and large game.[26] If so, we think there is a deep puzzle here, given the evolutionary timeline for know-how copying that was set out above. Should it turn out that big game hunting did not in fact evolve until somewhat later in our line, our arguments would stand in need of a corresponding temporal adjustment. However, the puzzle itself would remain, so long as the origins of big game hunting preceded the establishment of know-how copying in our line, a proposition we think is very plausible. Let us now explain.

A lifeway organized around big game hunting comes with imposing informational demands.[27] Hunters need to coordinate their behaviour at and over time; not just to succeed, but to drive down risks to manageable levels (hunting big game, say with basic stabbing spears, is about as dangerous as one can get!). In addition, regular hunting likely implies a division of labour, requiring coordination across the broader group (Wrangham 2017). Most hunts fail, and hunters need food to eat on days where hunts are unsuccessful (Kelly 2013). They cannot, in addition, spend hours targeting fallback foods or consuming low-quality foods when they come up empty handed. Moreover,

over insects, even when only tiny quantities of meat are expected, to gain essential micronutrients (see footnote 21) (Tennie et al. 2014).

[25] Whether *Homo erectus* hunted or power scavenged, they very likely were *obligate* tool manufacturers and users (Shea 2017); they *needed* their technology, especially to reliably gain access to, and remove, meat packages from carcasses. Additionally, *Homo erectus* may have relied upon their tools to transform large pieces of meat into smaller, more easily digestible pieces, by repeatedly cutting and pounding the meat (Zink and Lieberman 2016)—a scenario Tennie (2023) has called the 'steak tartare hypothesis'. The visible evidence from all these are the stone tools, as they survived in the record (though organic tools would likely have been in use as well).

[26] Consistent with this: Roach and colleagues have reported fossil footprint evidence dating to around 1.5 mya of what may well be an all-male *Homo erectus* hunting party patrolling a lakeside in East Africa (Callaway 2015).

[27] Hunting big or even medium-sized game, and *depending* on the spoils of such hunting, poses a range of complications that are absent from, e.g., chimp-style hunting of colobus monkeys.

9780192886491_Book.indb 565 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:05:02

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

hunters need to negotiate and control the distribution of profits of hunts in ways that do not undermine continued cooperation (Tomasello et al. 2012; Sterelny 2021; Planer 2023).[28] They also need to know things about the prey they target (e.g. *where* they move across the landscape, *when* they are most vulnerable). Plus, there are the informational demands of whatever technology they use to hunt and/or which other group members use to harvest and process fallback foods (though, importantly, as we have seen, these demands need not be know-how copying demands).[29]

To meet these informational demands, we think *Homo erectus* likely needed a communication system that was *richer* than that of modern great apes.[30] Not a full human language, to be sure, but something that was at least beginning to trend in the general direction of language, and from which language eventually evolved (we assume). In particular, we think they needed enhanced means for communicating the various 'w-' forms of information: know-when, know-where, etc. (though not (yet) know-how). But, and this is the puzzle: how could *Homo erectus* have managed to assemble and maintain a significantly enriched communication system, relative to a great-ape-*like* baseline, while still lacking know-how copying? As a first step towards solving this puzzle, it is helpful to distinguish two broad views about the origins of language. On one, language grew out of a communication system that was initially (predominately) gestural and iconic (see e.g. Donald 2001; Tomasello 2008; Arbib et al. 2008; Corballis 2011; Planer and Sterelny 2021). On the other, language was always vocal and generally arbitrary (see e.g. Wray 1998, 2002, 2005; Mithen 2006, 2009). These associations between gesture and iconicity, on the one hand, and vocalization and arbitrariness, on the other, are not coincidental. As has been pointed out many times (and as is evidenced by contemporary sign languages), the gesture–visual channel affords iconicity to a much greater extent than the vocal–auditory channel. We think our proposed timeline for know-how copying fits much better with the gestural–iconic view of language origins. For it is possible to understand how *Homo erectus* might have evolved a considerably richer communication system despite a lack of know-how copying *if* that system was mostly gestural and iconic. Not so, if that system was instead vocal and arbitrary.

Iconic signs represent things through *form resemblance*. An iconic sign for a spear, say, might take the form of a throwing motion (resembling a behaviour done with a spear), or of a pyramid shape made with the hands (resembling a part of the spear, its tip). A new, but intuitive distinction can be drawn between icons that resemble behaviour/artefact forms that lie *within* a focal individual's species typical ZLS, and icons that mimic forms that lie *beyond* this range. Producing chewing or sucking behaviour to request food would be an example of the former kind of icon for humans (and many other animals). *Pantomiming* the distinctive movements of a symphony conductor to mean classical

[28] Though the taking of prime prey can be expected to partially damp down this problem.

[29] As an example of a hunting technology that does not require know-how copying: consider the 'spears' which Fongoli chimpanzees fashion and use to hunt lesser bush babies (see Pruetz and Bertolani 2007).

[30] See Planer and Sterelny (2021) for an extended defence of this claim. See also Planer (2017a, b).

9780192886491_Book.indb 566 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:02

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

music, or copying the sinuous movement of a snake, would be examples of the latter. We take it that the production of icons of the first kind will also be elements of the relevant individual's ZLS. They are signs that can quite readily evolve via the gradual pathways of ritualization (in which a behaviour goes from cue to signal) or manipulation (in which a behaviour goes from a form of coercion to signal) (Maynard-Smith and Harper 2003; Scott-Phillips 2014).[31] But they are also signs that might realistically 'occur to' an individual of the species to 'try out' in communication with conspecifics, allowing for a novel iconic sign to be invented directly. They do not require know-how copying (and thus, are not examples of true panto*mim*ing). This may be part of the explanation for the origins of genital-involving signs in great apes that are used to request/initiate copulation.

In summary, there is no need to appeal to know-how copying to explain how *Homo erectus* might have evolved an expanded body (relative to other great apes) of gestural–iconic signs inasmuch as these signs recycled forms within their species' ZLS. On this view, the forms of these signs would have been individually invented by *Homo erectus* senders. However, to the extent that they converged on similar iconic forms, regularities would emerge. Two *Homo erectus* might independently come to use faux knapping behaviour with other individuals as a sign for suggesting that they all now go collect some stone. Receivers need not know or care about how the sign has come to exist to be appropriately informed by it. And from today's perspective, we need not postulate know-how copying for such signs.

Moving forwards: as we enter the Middle Pleistocene, hominin behaviour was becoming more different still. In particular, our ancestors evolved greater control over fire and began targeting a wider range of resources. This is attested to by sites such as Gersher Benot in Israel (dating to around 780 kya) (Goren-Inbar et al. 2000). They were also beginning to manufacture additional tools *and* more complex communication systems. It might be that a basic ability for know-how copying first evolved around this time, not least as such an ability would *then* have provided benefits. We finish this section by elucidating three general evolutionary scenarios:

1. *Know-how copying originally evolved for acquiring communicative forms (e.g. words or word-like elements) and was later co-opted for acquiring technological forms.* As hominin communication systems grew in complexity, individuals who could acquire signals from others via know-how copying would have been able to develop richer and more effective communicative repertoires, and would have expended less time and energy acquiring signals. In essence, they could piggyback on others' signal innovation efforts. It was only at some later point in time that know-how copying came to play a role in the technological domain.
2. *Know-how copying originally evolved for acquiring technological forms (e.g. certain stone tools) and was later co-opted for acquiring communicative forms.* Enhanced

[31] See Planer and Sterelny (2021: ch. 4) for an extended discussion of these and other pathways to signalhood.

9780192886491_Book.indb 567 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:05:02

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

types of ecological cooperation selected for more sophisticated technique and technology in our line. Individuals who could copy the know-how underlying these forms would have acquired these forms more rapidly and with less learning effort. It was only at some later point in time that know-how copying came to play a role in the communicative domain.[32]

3. *Know-how copying was always multi-functional in our line, servicing the acquisition of forms across both the above domains, and possibly others.* On this scenario, communication and technology co-evolved, and individuals who could engage in know-how copying would have been advantaged in both domains (for reasons already mentioned). Neither domain need be the primary 'home' of such copying. Selective pressure for know-how copying might have come from elsewhere, too. Certainly, that was true later on, as things like ritual, dance, and other conventional behaviours took on increasing importance for hominins. Note that, if this scenario is right, evidence for know-how copying in the technological domain doubles as evidence for know-how copying in the communicative domain. If true, once unambiguous signatures of know-how copying of technique and/or technology appear in the archaeological record, we could infer that the corresponding hominin tool makers were also capable of constructing large, expandable lexicons of non-iconic (i.e. arbitrary) signs.[33]

To the extent that early language was primarily gestural, the last scenario seems most likely. Individuals who (i) paid closer attention to others' manual behaviours, who (ii) were better at discerning the goals and intentions behind those behaviours, and who (iii) were better able to reproduce those behaviours—especially if these would even be beyond their ZLS—would have had an upper hand in acquiring both their group's communicative forms and technological forms. We incline towards this scenario, as we find the gestural origins view of language independently plausible.[34] Gesture is under voluntary control in other great apes, which is not surprising given their nature as extractive foragers (Genty et al. 2009). Moreover, how they gesture is shaped by their individual experiences. To a significant extent, great apes *learn* (though they do not learn via know-how copying) how to communicate with gestures. This includes how to tailor the form of a gesture depending on the attentional (and perhaps informational) state of the audience (Moore 2016). Gestures in them are also produced and interpreted in highly flexible ways that depend on context (Hobaiter and Byrne 2014). In these respects, great-ape gestural communication is already much more like human linguistic

[32] We think this is probably the default scenario of many theorists.

[33] For completeness' sake, we acknowledge that it is also possible that the evolution of know-how copying originally had nothing to do with either communication or technology. However, we find such a scenario highly implausible.

[34] See Planer and Sterelny (2021: especially, ch. 2). The following briefly summarizes the highlights. See also Planer and Reed (2021).

9780192886491_Book.indb 568 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:03

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

communication than is the vocal communication of other great apes.[35] The same would have been true of the earliest hominins. In addition, as already mentioned, gesture much more readily affords the use of iconic signs, and it is plausible that iconicity played a crucial role in allowing for effective communication in the absence of established communicative conventions. This chapter has presented yet another argument in support of a gestural origins view based on considerations about the evolutionary timeline of know-how copying in hominins. Namely, a gestural origins view alone enables us to see how, despite lacking abilities for know-how copying, early Pleistocene hominins were able to evolve the communicative upgrades that the archaeological evidence strongly suggests they evolved.

30.5 Summary

We began this chapter by setting out a novel social learning framework. The defining feature of this framework is that it foregrounds the type of information that is socially transmitted in a given case of culture, as opposed to esoteric social learning mechanisms on which a given case of culture is hypothesized to depend. Using this framework, we advanced a working hypothesis about the uniqueness of human cultural capacities among apes. According to this hypothesis, humans show a special propensity for the social transmission of supra-individual know-how, or in other words, for know-how copying. We then turned our attention to available methodologies that help us identify know-how copying. We pointed out that one of these, the method of local restriction, is especially well suited to helping us locate the origins of know-how copying in the hominin archaeological record. A thumbnail sketch was then provided of how the archaeological evidence—and in particular, of how evidence of hominin technological evolution—looks when viewed from the vantage point of this methodology. On the picture that emerges, the evolutionary origins of know-how copying are considerably more recent than has typically been assumed. Finally, we noted that this relatively late date for the evolution of know-how copying arguably poses a puzzle about the evolution of communication in our line (specifically, during the Mid-Early Pleistocene), which a gestural–iconic origins view of early language is uniquely poised to solve.

[35] Which is not to say that great-ape vocalization is completely automatic, or completely insensitive to experience, etc. But it is to say that, even given recent results showing more nuance and sophistication in vocal communication among other great apes than has traditionally been attributed to that domain (see e.g. Crockford et al. 2012, 2017; Schel et al. 2013), important contrasts remain between great-ape gestural and vocal communication along the above dimensions.

9780192886491_Book.indb 569 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:05:03

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Funding statement

This research received support from the DFG-Centre for Advanced Studies in the Humanities Words, Bones, Genes, Tools (DFG-KFG 2237), which is gratefully acknowledged.

Acknowledgements

We are grateful to Lauren Reed for helpful comments on an earlier draft.

References

Acerbi, A., Snyder, W. D., and Tennie, C. (2022). 'The method of exclusion (still) cannot identify specific mechanisms of cultural inheritance', *Scientific Reports* 12(1): 21680.

Arbib, M. A., K. Liebal, and S. Pika (2008). 'Primate vocalization, gesture, and the evolution of human language', *Current Anthropology* 49(6): 1053–1076.

Bandini, E. and Tennie, C. (2017). 'Spontaneous reoccurrence of "scooping", a wild tool-use behaviour, in naïve chimpanzees', *PeerJ* 5: e3814.

Bandini, E. and Tennie, C. (2018). 'Naive, captive long-tailed macaques (*Macaca fascicularis fascicularis*) fail to individually and socially learn pound-hammering, a tool-use behaviour', *Royal Society Open Science* 5(5): 171826.

Bandini, E., Motes-Rodrigo, A., Steele, M. P., et al. (2020). 'Examining the mechanisms underlying the acquisition of animal tool behaviour', *Biology Letters* 16(6): 20200122.

Bandini, E., Harrison, R. A., and Motes-Rodrigo, A. (2022). 'Examining the suitability of extant primates as models of hominin stone tool culture', *Humanities and Social Sciences Communications* 9(1): 1–18.

Boesch, C., Marchesi, P., Marchesi, N., et al. (1994). 'Is nut cracking in wild chimpanzees a cultural behaviour?', *Journal of Human Evolution* 26(4): 325–338.

Buskell, A. and Tennie, C. (2025). 'Mere recurrence and cumulative culture at the margins', *The British Journal for the Philosophy of Science* 76(1).

Byrne, R. W. (2007). 'Culture in great apes: Using intricate complexity in feeding skills to trace the evolutionary origin of human technical prowess', *Philosophical Transactions of the Royal Society B: Biological Sciences* 362: 577–585.

Callaway, E. (2015). 'Homo erectus footprints hint at ancient hunting party', *Nature News*, 17 April.

Chapais, B. (2008). *Primeval Kinship: How Pair-Bonding Gave Birth to Human Society*. Cambridge, MA: Harvard University Press.

Clay, Z. and Tennie, C. (2018). 'Is overimitation a uniquely human phenomenon? Insights from human children as compared to bonobos', *Child Development* 89(5): 1535–1544.

Conard, N. J. (2005). 'An overview of the patterns of behavioural change in Africa and Eurasia during the Middle and Late Pleistocene', in F. d'Errico, L. Backwell, and B. Malauzat (eds.), *From Tools to Symbols: From Early Hominids to Modern Humans*, 294–332. New York: NYU Press.

9780192886491_Book.indb 570 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:03

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Corballis, M. C. (2011). *The Recursive Mind: The Origins of Human Language, Thought, and Civilization*. Princeton: Princeton University Press.

Corbey, R., Jagich, A., Vaesen, K., and Collard, M. (2016). 'The Acheulean handaxe: More like a bird's song than a Beatles' tune?', *Evolutionary Anthropology: Issues, News, and Reviews* 25(1): 6–19.

Crockford, C., Wittig, R. M., et al. (2012). 'Wild chimpanzees inform ignorant group members of danger', *Current Biology* 22(24): 142–146.

Crockford, C., Wittig, R., et al. (2017). 'Vocalizing in chimpanzees is influenced by social-cognitive processes', *Science Advances* 3(11).

Davidson, I. (2016). 'Stone tools: Evidence of something in between culture and cumulative culture?', in M. N. Haidle, N. J. Conard, and M. Bolus (eds.), *The Nature of Culture: Based on an Interdisciplinary Symposium 'The Nature of Culture', Tübingen, Germany*, 99–120. Amsterdam: Springer.

Dennett, D. C. (1995). 'Darwin's dangerous idea', *The Sciences* 35(3): 34–40.

Dominguez-Rodrigo, M. and Pickering, T. R. (2017). 'The meat of the matter: An evolutionary perspective on human carnivory', *Azania: Archaeological Research in Africa* 52(1): 4–32.

Donald, M. (2001). *A Mind So Rare: The Evolution of Human Consciousness*. New York: W. W. Norton.

Fedato, A., Silva-Gago, M., Terradillos-Bernal, M., et al. (2020). 'Hand grasping and finger flexion during Lower Paleolithic stone tool ergonomic exploration', *Archaeological and Anthropological Sciences* 12: 1–9.

Finkel, M. and Barkai, R. (2018). 'The Acheulean handaxe technological persistence: A case of preferred cultural conservatism?', in *Proceedings of the Prehistoric Society* (Vol. 84), 1–19. Cambridge: Cambridge University Press.

Foley, R. (1987). 'Hominid species and stone-tool assemblages: How are they related?', *Antiquity* 61(233): 380–392.

Genty, E., Breuer, T., Hobaiter, C., and Byrne, R. W. (2009). 'Gestural communication of the gorilla (*Gorilla gorilla*): Repertoire, intentionality and possible origins', *Animal Cognition* 12: 527–546.

Goren-Inbar, N., Feibel, C. S., Verosub, K. L., et al. (2000). 'Pleistocene milestones on the out-of-Africa corridor at Gesher Benot Ya'aqov, Israel', *Science* 289(5481): 944–947.

Guthrie, R. D. (2007). 'Haak en Steek—The tool that allowed hominins to colonize the African savanna and flourish there', in W. Roebroeks (ed.), *Guts and Brains*, 133–164. Leiden: Leiden University Press.

Harmand, S., Lewis, J. E., Feibel, C. S., et al. (2015). '3.3-million-year-old stone tools from Lomekwi 3, West Turkana, Kenya', *Nature* 521(7552): 310–315.

Henshilwood, C. S., d'Errico, F., Yates, R., et al. (2002). 'Emergence of modern human behavior: Middle Stone Age engravings from South Africa', *Science* 295(5558): 1278–1280.

Heyes, C. (2018). *Cognitive Gadgets: The Cultural Evolution of Thinking*. Cambridge, MA: Harvard University Press.

Hobaiter, C. and Byrne, R. W. (2014). 'The meanings of chimpanzee gestures', *Current Biology* 24(14): 1596–1600.

Hrdy, S. B. (2009). *Mothers and Others: The Evolutionary Origins of Mutual Understanding*. Cambridge, MA: Harvard University Press.

Isaac, G. L. and Isaac, B. (1977). *Olorgesailie: Archaeological Studies of a Middle Pleistocene Lake Basin*. Chicago, IL: University of Chicago Press.

9780192886491_Book.indb 571 /12_first_proofs/first_pr[illegible]/xml_for_typesetting 27-Mar-25 23:05:03

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Kelly, R. L. (2013). *The Lifeways of Hunter-Gatherers: The Foraging Spectrum*. Cambridge: Cambridge University Press.

Key, A. J. and Lycett, S. J. (2019). 'Biometric variables predict stone tool functional performance more effectively than tool-form attributes: A case study in handaxe loading capabilities', *Archaeometry* 61(3): 539–555.

Koops, K., Soumah, A. G., van Leeuwen, K. L., et al. (2022). 'Field experiments find no evidence that chimpanzee nut cracking can be independently innovated', *Nature Human Behaviour* 6(4): 487–494.

Kuhn, S. L. (2020). *The Evolution of Paleolithic Technologies*. Abingdon: Routledge.

Laland, K., Matthews, B., and Feldman, M. W. (2016). 'An introduction to niche construction theory', *Evolutionary Ecology* 30: 191–202.

Laland, K. N. and Janik, V. M. (2006). 'The animal cultures debate', *Trends in Ecology & Evolution* 21(10): 542–547.

Lesnik, J. J. (2018). *Edible Insects and Human Evolution*. Gainesville: University of Florida Press.

Lombard, M. (2012). 'Thinking through the Middle Stone Age of sub-Saharan Africa', *Quaternary International* 270: 140–155.

Lombard, M. and Shea, J. J. (2021). 'Did Pleistocene Africans use the spearthrower-and-dart?', *Evolutionary Anthropology: Issues, News, and Reviews* 30(5): 307–315.

Ludwig, B. V. and Harris, J. W. K. (1998). 'Towards a technological reassessment of East African Plio-Pleistocene lithic assemblages', *Early Human Behaviour in Global Context: The Rise and Diversity of the Lower Palaeolithic Record* 84: 107.

Maslin, M. (2016). *The Cradle of Humanity: How the Changing Landscape of Africa Made Us So Smart*. Oxford: Oxford University Press.

McPherron, S. P., Alemseged, Z., Marean, C. W., et al. (2010). 'Evidence for stone-tool-assisted consumption of animal tissues before 3.39 million years ago at Dikika, Ethiopia', *Nature* 466(7308): 857–860.

Milton, K. (1999). 'A hypothesis to explain the role of meat-eating in human evolution', *Evolutionary Anthropology Issues News and Reviews* 8(1): 11–21.

Mithen, S. (2009). 'Holistic communication and the coevolution of language and music: Resurrecting an old idea', *The Prehistory of Language* 58.

Mithen, S. J. (2006). *The Singing Neanderthals: The Origins of Music, Language, Mind, and Body*. Cambridge, MA: Harvard University Press.

Moore, M. W. and Perston, Y. (2016). 'Experimental insights into the cognitive significance of early stone tools', *PLoS One* 11(7): e0158803.

Moore, R. (2016). 'Meaning and ostension in great ape gestural communication', *Animal Cognition* 19(1): 223–231.

Motes-Rodrigo, A. and Tennie, C. (2021). 'The method of local restriction: In search of potential great ape culture-dependent forms', *Biological Reviews* 96(4): 1441–1461.

Motes-Rodrigo, A., McPherron, S. P., Archer, W., et al. (2022). 'Experimental investigation of orangutans' lithic percussive and sharp stone tool behaviours', *PLoS One* 17(2): e0263343

Nakahashi, W., Horiuchi, S., and Ihara, Y. (2018). 'Estimating hominid life history: The critical interbirth interval', *Population Ecology* 60: 127–142.

Neadle, D. L., Chappell, J., Clay, Z., and Tennie, C. (2021). 'Even under majority influence, great apes fail to copy novel actions', OSF preprint, https://doi.org/10.31219/osf.io/swt9b.

Neadle, D., Allritz, M., and Tennie, C. (2017). 'Food cleaning in gorillas: Social learning is a possibility but not a necessity', *PLoS One* 12(12): e0188866.

9780192886491_Book.indb 572 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:03

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Opie, K. and Power, C. (2009). 'Grandmothering and female coalitions', *Early Human Kinship: From Sex to Social Reproduction* 168.

Pickering, T. R. and H. Bunn (2012). 'Meat foraging by Pleistocene African hominins: Tracking behavioral evolution beyond baseline inferences of early access to carcasses', in M. Dominguez-Rodrigo (ed.), *Stone Tools and Fossil Bones*, 152–173. New York: Cambridge University Press.

Pilbeam, D. R., Lieberman, D. E., Muller, M. N., and Wrangham, R. W. (2017). 'Reconstructing the last common ancestor of chimpanzees and humans', in M. N. Muller, R. W. Wrangham, and D. R. Pilbeam (eds.), *Chimpanzees and Human Evolution*, 22–141. Cambridge, MA: Harvard University Press.

Planer, R. J. (2017a). 'Protolanguage might have evolved before ostensive communication', *Biological Theory* 12(2): 72–84.

Planer, R. J. (2017b). 'Talking about tools: Did Early Pleistocene hominins have a protolanguage?', *Biological Theory* 12(4): 211–221.

Planer, R. J. and Reed, L. W. (2021). 'Like hand, like mouth: On the role of gesture-linked mouth actions in the evolution of language', *Biological Theory* 16: 90–101.

Planer, R. and Sterelny, K. (2021). *From Signal to Symbol: The Evolution of Language*. Cambridge, MA: MIT Press.

Planer, R. J. (2023). 'Cooperative breeding and the evolutionary origins of shared intentionality', *Philosophy, Theory, and Practice in Biology* 15.

Pope, S. M., Taglialatela, J. P., Skiba, S. A., and Hopkins, W. D. (2018). 'Changes in frontoparietotemporal connectivity following do-as-I-do imitation training in chimpanzees (*Pan troglodytes*)', *Journal of Cognitive Neuroscience* 30(3): 421–431.

Proffitt, T., Luncz, L. V., Falótico, T., et al. (2016). 'Wild monkeys flake stone tools', *Nature* 539(7627): 85–88.

Pruetz, J. D. and Bertolani, P. (2007). 'Savanna chimpanzees, Pan troglodytes verus, hunt with tools', *Current Biology* 17(5): 412–417.

Reindl, E., Bandini, E., and Tennie, C. (2018). 'The zone of latent solutions and its relation to the classics: Vygotsky and Köhler', *Evolution of Primate Social Cognition* 231–248.

Richerson, P. J. and Boyd, R. (2008). *Not by Genes Alone: How Culture Transformed Human Evolution*. Chicago, IL: University of Chicago Press.

Richerson, P. J. and Boyd, R. (2020). 'The human life history is adapted to exploit the adaptive advantages of culture', *Philosophical Transactions of the Royal Society B* 375(1803): 20190498.

Rizal, Y., Westaway, K. E., Zaim, Y., et al. (2020). 'Last appearance of Homo erectus at Ngandong, Java, 117,000–108,000 years ago', *Nature* 577(7790): 381–385.

Roach, N. T., Hatala, K. G., Ostrofsky, K. R., et al. (2016). 'Pleistocene footprints show intensive use of lake margin habitats by Homo erectus groups', *Scientific Reports* 6(1): 26374.

Robson, S. L. and Wood, B. (2008). 'Hominin life history: Reconstruction and evolution', *Journal of Anatomy* 212(4): 394–425.

Rolian, C. and Carvalho, S. (2017). 'Tool use and manufacture in the last common ancestor of Pan and Homo', in M. N. Muller, R. W. Wrangham, and D. R. Pilbeam (eds.), *Chimpanzees and Human Evolution*, 602–644. Cambridge, MA: The Belknap Press of Harvard University Press.

Rothman, J. M., Raubenheimer, D., Bryer, M. A., et al. (2014). 'Nutritional contributions of insects to primate diets: Implications for primate evolution', *Journal of Human Evolution* 71: 59–69.

9780192886491_Book.indb 573 /12_first_proofs/first_pr/xml_for_typesetting 27-Mar-25 23:05:03

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Sahle, Y. and Brooks, A. S. (2019). 'Assessment of complex projectiles in the early Late Pleistocene at Aduma, Ethiopia', *PLoS One* 14(5): e0216716.

Sanz, C., Call, J., and Morgan, D. (2009). 'Design complexity in termite-fishing tools of chimpanzees (*Pan troglodytes*)', *Biology Letters* 5(3): 293–296.

Schel, A. M., Townsend, S. W., et al. (2013). 'Chimpanzee alarm call production', in M. Camps and P. Chauhan (eds.), *Sourcebook of Paleolithic Transitions*. New York: Springer.

Scott-Phillips, T. (2014). *Speaking Our Minds: Why Human Communication Is Different, and How Language Evolved to Make It Special*. London: Bloomsbury Publishing.

Semaw, S., Rogers, M., and Stout, D. (2009). 'The Oldowan-Acheulian transition: Is there a "Developed Oldowan" artifact tradition?', *Sourcebook of Paleolithic Transitions: Methods, Theories, and Interpretations* 173–193.

Shea, J. J. (2009). 'The impact of projectile weaponry on Late Pleistocene hominin evolution', *The Evolution of Hominin Diets: Integrating Approaches to the Study of Palaeolithic Subsistence*, 189–199. New York: Springer.

Shea, J. J. (2017). 'Occasional, obligatory, and habitual stone tool use in hominin evolution', *Evolutionary Anthropology: Issues, News, and Reviews* 26(5): 200–217.

Smith, J. M. and Harper, D. (2003). *Animal Signals*. Oxford: Oxford University Press.

Snyder, W. D., Reeves, J. S., and Tennie, C. (2022). 'Early knapping techniques do not necessitate cultural transmission', *Science Advances* 8(27): eabo2894.

Speth, J. D. (2017). 'Putrid meat and fish in the Eurasian middle and upper Paleolithic: Are we missing a key part of Neanderthal and modern human diet', *PaleoAnthropology* 2017: 44–72.

Speth, J. D. and Morin, E. (2022). 'Putrid meat in the Tropics: It wasn't just for Inuit', *PaleoAnthropology* 2022(2).

Sterelny, K. (2012). *The Evolved Apprentice*. Cambridge, MA: MIT Press.

Sterelny, K. (2021). *The Pleistocene Social Contract: Culture and Cooperation in Human Evolution*. Oxford: Oxford University Press.

Stout, D., Semaw, S., Rogers, M. J., and Cauche, D. (2010). 'Technological variation in the earliest Oldowan from Gona, Afar, Ethiopia', *Journal of Human Evolution* 58(6): 474–491.

Tennie, C. (2019). 'Could nonhuman great apes also have cultural evolutionary psychology?', *Behavioral and Brain Sciences* 42: 37–39.

Tennie, C. (2023). 'The earliest tools and cultures of hominins', in J. J. Tehrani, J. Kendal, and R. Kendal (eds.), *Oxford Handbook of Cultural Evolution*. Oxford: Oxford University Press.

Tennie, C. and Call, J. (2023). 'Unmotivated subjects cannot provide interpretable data and tasks with sensitive learning periods require appropriately aged subjects: A commentary on Koops et al. (2022) "Field experiments find no evidence that chimpanzee nut cracking can be independently innovated" ', *Animal Behavior and Cognition* 10(1): 89–94.

Tennie, C., Call, J., and Tomasello, M. (2009). 'Ratcheting up the ratchet: On the evolution of cumulative culture', *Philosophical Transactions of the Royal Society B: Biological Sciences* 364(1528): 2405–2415.

Tennie, C., Gilby, I. C., and Mundry, R. (2009). 'The meat-scrap hypothesis: Small quantities of meat may promote cooperative hunting in wild chimpanzees (*Pan troglodytes*)', *Behavioral Ecology and Sociobiology* 63: 421–431.

Tennie, C., Call, J., and Tomasello, M. (2012). 'Untrained chimpanzees (*Pan troglodytes schweinfurthii*) fail to imitate novel actions', *PLoS ONE* 7: e41548.

9780192886491_Book.indb 574 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:03

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Tennie, C., O'Malley, R. C., and Gilby, I. C. (2014). 'Why do chimpanzees hunt? Considering the benefits and costs of acquiring and consuming vertebrate versus invertebrate prey', *Journal of Human Evolution* 71: 38–45.

Tennie, C., Braun, D. R., Premo, L. S., and McPherron, S. P. (2016). 'The Island test for cumulative culture in the Paleolithic', in M. N. Haidle, N. J. Conard, and M. Bolus (eds.), *The Nature of Culture: Based on an Interdisciplinary Symposium 'The Nature of Culture', Tübingen, Germany*, 121–133. Amsterdam: Springer.

Tennie, C., Bandini, E., Van Schaik, C. P., and Hopper, L. M. (2020). 'The zone of latent solutions and its relevance to understanding ape cultures', *Biology & Philosophy* 35: 1–42.

Texier, P. J., Porraz, G., Parkington, J., et al. (2013). 'The context, form and significance of the MSA engraved ostrich eggshell collection from Diepkloof Rock Shelter, Western Cape, South Africa', *Journal of Archaeological Science* 40(9): 3412–3431.

Thieme, H. (1997). 'Lower Palaeolithic hunting spears from Germany', *Nature* 385(6619): 807–810.

Tomasello, M. (1999). 'The human adaptation for culture', *Annual Review of Anthropology* 28(1): 509–529.

Tomasello, M. (2008). *Origins of Human Communication*. Cambridge, MA: MIT Press.

Tomasello, M., Call, J., and Gluckman, A. (1997). 'Comprehension of novel communicative signs by apes and human children', *Child Development* 68: 1067–1080.

Tomasello, M., Melis, A. P., Tennie, C., et al. (2012). 'Two key steps in the evolution of human cooperation: The interdependence hypothesis', *Current Anthropology* 53(6): 673–692.

Toth, N. and Schick, K. (2019). 'Why did the Acheulean happen? Experimental studies into the manufacture and function of Acheulean artifacts', *L'Anthropologie* 123(4–5): 724–768.

van Schaik, C. P. (2010). 'Social learning and culture in animals', *Animal Behaviour: Evolution and Mechanisms* 623–653.

Van Schaik, C. P. and Burkart, J. M. (2010). 'Mind the gap: Cooperative breeding and the evolution of our unique features', in Peter M. Kappeler and Joan Silk (eds.), *Mind the Gap: Tracing the Origins of Human Universals* 477–496. New York: Springer.

Wadley, L. (2015). 'Those marvellous millennia: The Middle Stone Age of southern Africa', *Azania: Archaeological Research in Africa* 50(2): 155–226.

Westergaard, G. C. and Suomi, S. J. (1994). 'A simple stone-tool technology in monkeys', *Journal of Human Evolution* 27(5): 399–404.

Whiten, A., Goodall, J., McGrew, W. C., et al. (1999). 'Cultures in chimpanzees', *Nature* 399(6737): 682–685.

Whiten, A., Goodall, J., McGrew, W. C., et al. (2001). 'Charting cultural variation in chimpanzees', *Behaviour* 138: 1481–1516.

Wilkins, J. and M. Chazan (2012). 'Blade production 500 thousand years ago at Kathu Pan 1, South Africa: Support for a multiple origins hypothesis for Early Middle Pleistocene blade technologies', *Journal of Archaeological Science* 39: 1883–1900.

Wilkins, J., Schoville, B. J., et al. (2012). 'Evidence for early hafted hunting technology', *Science* 338: 942–945.

Wrangham, R. (2017). 'Control of fire in the Paleolithic: Evaluating the cooking hypothesis', *Current Anthropology* 58(S16): S303–S313.

Wray, A. (1998). 'Protolanguage as a holistic system for social interaction', *Language and Communication* 18(1): 47–67.

9780192886491_Book.indb 575 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:05:03

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Wray, A. (2002). 'Dual processing in protolanguage: performance without competence', in A. Wray (ed.), *The Transition to Language*, 113–137. Oxford: Oxford University Press.

Wray, A. (2005). 'The explanatory advantages of the holistic protolanguage model: The case of linguistic irregularity', *Behavioral and Brain Sciences* 28(2): 147–148.

Zhang, P., Huang, W., and Wang, W. (2010). 'Acheulean handaxes from Fengshudao, Bose sites of south China', *Quaternary International* 223: 440–443.

Zink, K. D. and Lieberman, D. E. (2016). 'Impact of meat and Lower Palaeolithic food processing techniques on chewing in humans', *Nature* 531(7595): 500–503.

9780192886491_Book.indb 576 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:03

CHAPTER 31

COMPARATIVE NEUROIMAGING TO STUDY THE NEURAL INFRASTRUCTURE FOR LANGUAGE

NICOLE EICHERT AND ROGIER B. MARS

31.1 Introduction

Understanding language evolution requires understanding the evolution of the neural machinery that enables this complex behaviour. Unfortunately, the brain is soft tissue which generally does not fossilize. Although endocasts can provide some information about the size and shape of our ancestors' brains, these measures are by necessity limited in the anatomical detail they can provide. For instance, endocasts lack information about the relative size of some of the brain's constituent parts, which could provide critical insights into the constraints placed on brain evolution (Bruner 2019). An alternative approach is to compare the human brain to that of other living species. As the evolutionary relationships between living mammalian species are now largely, although not completely (Foley et al. 2016), understood, comparing their brain architecture in a phylogenetic framework has the potential to elucidate both shared and unique aspects of human brain organization and how this relates to language abilities (Mars, Eichert, et al. 2018).

Empirical frameworks to characterize species differences in neurobiology have traditionally followed two separate approaches. The first one is to correlate the relationship between simplified descriptors of the brain with diverse behaviours in a large number of species. This approach exploits the fact that measures of brain size or the relative size of large parts of the brain are relatively easy to access post-mortem and

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

have been described in extensive databases (Brodmann 1909; Stephan et al. 1981; Elston 2007). A prominent example is the work showing a relationship between the size of the neocortex in different primate species and the complexity of their social network as indicated by their average group size (Dunbar 2003). The major disadvantage of this approach, however, is that measures such as brain size are often an aggregate measure of features at lower levels of brain organization, such as connectivity and arealization. As a result, many different aspects of cognition have been correlated with brain size and overall the conclusions have been unspecific (Healy 2021).

The second approach is to examine a small number of species, typically two, and to characterize neurobiological differences in a particular brain structure with high anatomical precision. The vocal learning circuit, for example, has been mapped in great detail in songbirds and compared to non-songbirds (Jarvis 2004) and similar comparisons have been made between humans and macaques (Jarvis 2019). However, the invasive and laborious nature of this work makes it unfeasible, and in cases of our closest animal relatives, unethical and illegal, to perform this work on a large scale across many species.

As these examples demonstrate, comparative neuroscience work traditionally has had either great breadth in terms of the numbers of species studied or great depth in terms of its anatomical detail, but rarely both. We argue that the availability of new types of data and the subsequent development of new analysis frameworks now allow us to combine the strengths of the two approaches. New developments in comparative neuroimaging have paved the way to study a large number of species with high spatial resolution at the whole-brain level (see Mars et al. 2014 and Friedrich et al. 2020 for reviews). In this chapter, we highlight some of the new data resources, discuss the new analysis frameworks, and review new studies on the neurobiology of language from the perspective of primate comparative neuroscience. We discuss how new data and new analysis approaches can enhance our understanding of language evolution, demonstrating how specializations of the human brain can be better understood in the context of a range of primate brains that allows a wider phylogenetic perspective.

31.2 New developments in data availability and data infrastructure

The first major development towards large-scale comparative neuroscience is the increasing availability of high-throughput methods for collecting anatomical data (Figure 31.1A). Neuroimaging techniques, in particular magnetic resonance imaging (MRI), have been the most prominent. However, more traditional anatomical techniques, such as histological or receptor-based mapping, are also becoming increasingly suitable for high-throughput approaches. New techniques for mapping gene expression, finally, provide completely new datatypes that are invaluable to comparative studies. We will discuss each of these developments in turn.

9780192886491_Book.indb 578 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:03

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

MRI was established in the early 1980s and rapidly emerged as the key method for computational neuroanatomy. MRI scanners generate electromagnetic waves and measure MR signals emitted from hydrogen atoms. By spatially varying the magnetic gradients, we can reconstruct a three-dimensional (3D) image of an object. The MR signal is obtained at discrete samples in space, i.e. volumetric pixels or voxels. Spatial image contrast in an MRI scan arises from variations in magnetic susceptibility of the tissue, which is governed by the macromolecular constitution of the specimen. The settings and timings of the scanner, i.e. the scanning sequence, determine the sensitivity of the MR signal towards a certain biological tissue property. Different sequences therefore allow one to study different aspects of the tissue (Lerch et al. 2017), including the presence of white matter fibres, the myelin and iron contents of the tissue, and the orientation of neurites. For example, in a conventional anatomical brain scan, such as the T1-weighted (T1w) scan, the cerebral cortex and deep brain nuclei—the 'grey matter' of the brain—appear darker than the remaining tissue of brain connections—the 'white matter' of the brain.

Within the human language literature, the advantages of combing different sequences is nicely illustrated by a recent study by Eichert and colleagues (2020). The authors wanted to find out whether the two laryngeal motor representations in the human brain were located in different brain areas. They therefore used functional MRI to locate the laryngeal representations in the brain and combined this with multiple parametric mapping (MPM; Weiskopf et al. 2013) in the same subjects. MPM uses a series of MRI sequences sensitive to different aspects of cortical microstructure, for example myelin content. Primary motor areas have higher myelin content than premotor areas. The authors then showed that the dorsal larynx representation has values similar to those of primary areas, whereas a more ventral representation bore hallmarks of a premotor region. This result demonstrates that the two human larynx representations are in anatomically distinct brain areas.

The non-invasive nature of MRI makes it highly suitable for comparative studies. Early work used T1w sequences to obtain whole-brain data from a range of species, allowing them to measure the relative size of parts of the brain at large scale across species (Rilling and Insel 1998; Rilling and Seligman 2002). Quite soon, functional MRI in anaesthetized animals followed, providing a more detailed cross-species comparison of the organization of specific brain networks (Vincent et al. 2007; Margulies et al. 2009; Mars et al. 2011). Such studies have now been performed on primates, carnivores, and rodents (see Mars et al. 2014 for a review). To obtain the breadth of the earlier large-scale comparative work, however, requires data to be acquired from post-mortem samples. If brain samples are processed quickly enough after death, fixative methods can preserve most structural tissue properties to a level that is comparable with the *in-vivo* organ. Diffusion MRI sensitive to the presence and directions of white matter pathways has been particularly successful in obtaining large post-mortem datasets (Assaf et al. 2020; Bryant et al. 2021).

Notwithstanding these successes, obtaining high-quality data in many species still presents a major challenge. The samples are often difficult to obtain, especially in the case of rare and long-living primates; protocols for tissue preparation are elaborate, the

9780192886491_Book.indb 579 /12_first_proofs/first_pr…/xml_for_typesetting 27-Mar-25 23:05:03

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

scanning sequences are specialized, and the hardware can be non-standard. Many of these challenges are overcome by collaborative efforts to construct shared databases. Tissue samples of primates are made available through repositories such as the Primate Brain Bank (Kaas and Van Eden 2011). In addition, MRI data resulting from scanning non-human species *in-vivo* or ***ex-vivo*** are increasingly made openly available online. An excellent example is the PRIMatE Data Exchange (PRIME-DE) platform, which is an open-science resource for the neuroimaging community giving access to non-human primate MRI data (Milham et al. 2018). Another example is the Digital Brain Zoo, part of the Oxford's Digital Brain Bank for post-mortem imaging data, which provides open access to imaging and microscopy data from multiple species (Tendler et al. 2022). Further advances in the data infrastructure relate to the standardization of data collection by adapting acquisition and preprocessing standards originally developed for human MRI big data projects, such as the Human Connectome Project (Autio et al. 2020). An overview of some recent resources is given in (Thiebaut de Schotten et al. 2019).

Whilst these developments greatly increased the sample sizes and species available for comparative mapping of the brain, in parallel we observed strong advances in high-throughput techniques to characterize the brain at the microscopic level. One example is the BigBrain database, which openly provides a post-mortem 3D histological reconstruction of a human brain combined with MRI data in the same individual (Amunts et al. 2013). The highly detailed multimodal atlas serves as a microscopic reference template for the human brain and allows us to link gold-standard anatomical investigations with macroscopic MRI. Human neurotransmitter receptor maps are currently not available at whole-brain coverage, but language-relevant regions have extensively been mapped using autoradiography (Amunts et al. 2010). In the macaque, however, whole-brain data for neurotransmitter receptor organization are available together with MRI data of the same individual (Froudist-Walsh et al. 2023). Using MRI as a bridging modality, such datasets can now be integrated across species to study microanatomical differences.

Another example of anatomical 'deep mapping' is the BigMac Database (Howard et al. 2023). This openly accessible resource integrates *in-vivo* and extensive post-mortem MRI data with multiple high-resolution histological contrasts in a single whole macaque brain. It provides not only a basis to validate MRI-based estimates of microstructure across spatial scales, but also to bridge to other species, where data is available at different levels. A recent comparison of BigBrain and BigMac, for example was able to link cross-species conservation in hippocampal microstructure with reorganization in functional connectivity (Eichert et al. 2024).

Another recent development worth mentioning is the increasing availability of spatial transcriptomics (Keil et al. 2018). Although the technical requirements to do this on a large scale are still out of reach for most labs, publicly available datasets are again proving very beneficial to the field at last. Prime examples are the Allen Human Brain atlas and the Allen Mouse Brain atlas (Lein et al. 2006; Jones et al. 2009). Taken together, the increased quality, availability, and coverage

9780192886491_Book.indb 580 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:03

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

of multimodal data from both human and non-human brains has made it possible to generate new testable hypothesis about the evolution of the neural infrastructure supporting language.

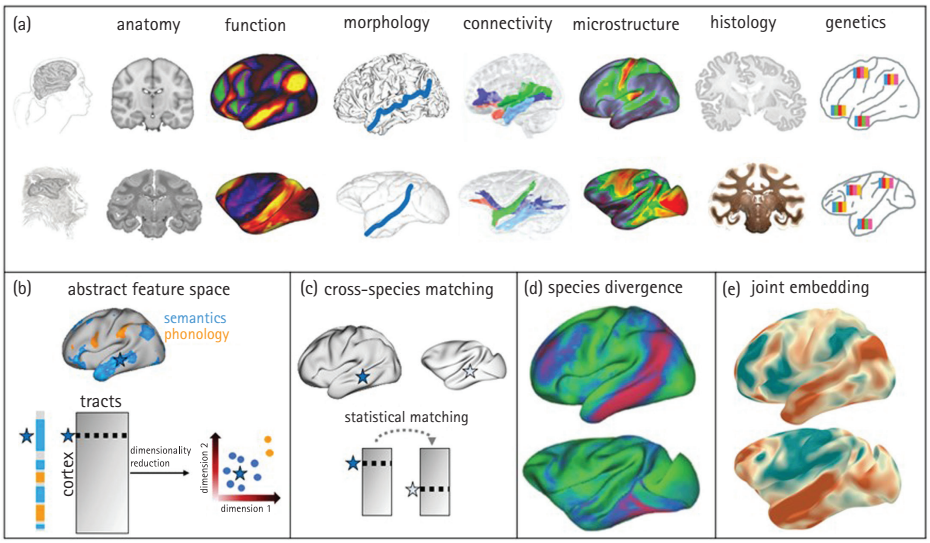


FIGURE 31.1 New developments in comparative neuroimaging.

A: Data of various descriptors of the brain are available across multiple species at comparable quality. These can be used to define common anchors for cross-species comparisons. Human and macaque brain are not to scale. **B**: Spatial maps of the brain can be represented in abstract matrix form, or features space, to which mathematical manipulations, such as dimensionality reduction, can be applied. **C**: Building shared feature spaces allows us to apply statistical methods to study shared and divergent aspects of the brain. **D**: Species divergence based on tractography reveals the highest divergence across human and macaque in the parietal and the temporal lobe. **E**: Joint species embedding can uncover spatial gradients that map comparable parts of the brain.

Sources: a) Human anatomy: The MNI152 human average brain template. © 1993–2009 Louis Collins, McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University; Macaque anatomy, morphology, histology: based on Howard et al. (2023), adapted with permission; Human and macaque function: adapted from Hayashi et al. (2021), under a CC BY 4.0 license; Human morphology: data from Eichert et al. (2020); Human and macaque connectivity: adapted from Roumazeilles et al. (2020), under a CC BY 4.0 license; Human and macaque microstructure: adapted from Eichert et al. (2020), under a CC BY 4.0 license; Human histology: adapted from Paquola et al. (2021), under a CC BY 4.0 license. d) Adapted from Mars, Eichert, et al. (2018), under a CC BY 4.0 license. e) Adapted from Eichert et al. (2024), under a CC BY-NC 4.0 license. All other images created by the authors.

31.3 The emergence of new analysis frameworks

Early comparative work using neuroimaging essentially replicated traditional anatomical approaches, such as measuring (partial) brain volumes (Rilling and Insel 1998) or

9780192886491_Book.indb 581 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:05:04

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

the extent of individual white matter tracts (Rilling et al. 2008). We gained invaluable anatomical insights from these cross-species studies, and, from a technical point of view, the careful comparison of traditional and novel methodology helped to recognize and mitigate limitations in either approach (Campbell and Pike 2014; Reveley et al. 2015). However, the field began to fully leverage its potential only when new analysis frameworks were designed to exploit the specific characteristics of the vast amounts and types of newly available data.

With this in mind, we have developed what we call the 'common space approach'. The approach rests on three pillars that exploit aspects of the newly available data (Mars et al. 2021). First, we exploit the multimodal nature of the data. As illustrated above, data quantifying multiple aspects of brain organization can be collected from the same brain sample non-destructively and repeatedly using neuroimaging. This allows us to move beyond the traditional limitation of being able to stain any part of the brain only for very few specific anatomical details. Second, we exploit the fact that MRI gives us information at high resolution in 3D, meaning that we do not just have summary information about whole brains or one slice, but instead detailed information covering the entire brain at millimetre or submillimetre resolution. Third, we exploit the digital and manipulable nature of the data to apply computational techniques that allows us to describe the brains in terms of organizational features that can be meaningfully compared between brains even if they differ substantially in size and morphology. In essence, we can now abstract away from the physical space or confounding features of the brain to focus only on those aspects that are relevant to the investigation at hand.

It is the last aspect of our framework—describing the brains in terms of an abstract feature space—that presents the most radical change compared to traditional comparative neuroscience. The idea is perhaps best illustrated in the domain of brain connectivity. Let us assume we compare brain activations during two language tasks (Figure 31.1B). In terms of the anatomical location of activation peaks in 3D physical space, we find a distributed pattern for each task without any obvious clusters. However, if we construct an abstract 'connectivity space', where regions with a similar profile of connections are located close together, we can uncover more meaningful relationships. In such a connectivity space, the peaks of one task will be located closer to one another than to the peaks of the other task; activations of the two tasks form two obvious clusters. The representation of activation peaks in terms of the abstract—in this case connectivity—space is clearly a more meaningful way of describing them than in terms of physical space (Mars, Passingham, et al. 2018). It should be mentioned that in many situations, the original feature space has many dimensions, or features, which contain redundant information and are difficult to visualize. Low-dimensional data embedding techniques are therefore often employed to derive a few principal axes or components for ease of interpretation (Mars, Passingham, et al. 2018; Xu et al. 2020; Eichert et al. 2024).

The critical advantage of using an abstract feature space is that the embedding technique can also be employed to compare brains (Figure 31.1C). As long as the feature space is shared across the brains under comparison, one can describe any part of the brain in the same abstract space. We refer to this concept therefore as the 'common

9780192886491_Book.indb 582 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:04

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

space approach' for comparative neuroscience. To illustrate this argument, we will stick with the example of brain connectivity. We know that the general template, or blueprint, of long-range brain connections is shared across mammals. Leveraging the advances of diffusion MRI, we can therefore reconstruct the path of homologous white matter tracts in different species. This set of tracts can then serve as the basis, or features, for a common space in which to represent the brain.

As a simple demonstration, we can pick an activation peak in the human brain, measure the connectivity strength with all tracts, and find the 'best match' for this connectivity profile in another species' brain. Thereby, we are essentially translating between species and predicting where an activation peak would be in another species. Expanding this principle to the whole brain instead of a single peak means that we can project or predict a human brain map in another species and vice versa. Importantly, not all points in the human brain will have a good match in other species and a brain map of similarity or divergence is on its own informative about human brain specializations (Figure 31.1D). Using this principle we were able to show that the human temporal lobe has undergone most significant remodelling in terms of structural connections compared to the macaque monkey (Mars, Sotiropoulos, et al. 2018). The flip-side of this approach is that we can also reconstruct spatial axes of variation that are shared across species using joint-embedding frameworks (Figure 31.1E; Eichert et al. 2024).

Importantly, the common space does not necessarily have to be based on connectivity. For instance, Beauchamp et al. (2022) recently used homologous genes to construct a common space, comparing mouse and human brains based on spatial patterns of gene expression. Establishing homologies reliably across species is not trivial and ultimately requires an iterative approach comparing shared and divergent features of anatomy. Any meaningful aspect of brain organization can potentially serve as the basis of the common space. Other examples include macroscopic anatomical landmarks (van Essen and Dierker 2007; Hill et al. 2010; Eichert et al. 2019; Heuer et al. 2019), cytoarchitecture (Petrides 1999; Zilles and Amunts 2010), cortical microstructure (Glasser et al. 2012), and brain activation (Sallet et al. 2011; Neubert et al. 2014; Xu et al. 2020; Balezeau et al. 2020).

The common space approach is especially powerful when applied to data across multiple modalities. A system as complex as language relies on adaptations on many anatomical levels and evolution acted differentially on different aspects of the brain's architecture (Mars, Eichert, et al. 2018). We can only really appreciate the neurobiological changes involved when we are able to dissociate how different aspects of neurobiology have changed relative to each other. Using the common space approach we can do precisely that, i.e. transform brains of multiple species to a reference space based on one modality and then to assess how other aspects changed relative to this transformation. The following section will briefly review some recent insights that we gained regarding language-relevant brain adaptations. A graphical summary is shown in Figure 31.2. The section is, of course, not a comprehensive review of the primate neurobiology of language and we focus on a selection of studies that leveraged a common

9780192886491_Book.indb 583 /12_first_proofs/first_pr[illegible]/xml_for_typesetting 27-Mar-25 23:05:04

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

space approach for a species comparison. For a more exhaustive literature overview, the reader is referred to reviews such as Friederici (2017).

31.4 What have we learned about the neural infrastructure of language?

One of the most noteworthy advancements in the field of comparative neuroimaging was to re-examine the expansion of the arcuate fasciculus (AF), a key pathway supporting language function in humans. Together with parts of the superior longitudinal fasciculus, the AF forms the 'dorsal' language pathway in the brain. The course of the human AF has first been studied comparatively using neuroimaging more than a decade ago (Rilling et al. 2008). It was observed that the human AF extends far into the temporal lobe, whilst the temporal extension is less prominent in great apes and barely present in macaque monkeys. This extension was thought to reflect a significant evolutionary modification of the course of the AF allowing it to connect new areas in the human lobe (Rilling et al. 2012). It was already known that the area of the brain that the posterior part of the AF connects to had expanded in the human since their last common ancestor with the macaque (van Essen and Dierker 2007) and this could potentially explain the comparative observations. However, at this time, this relationship between brain expansion and AF extension remained an implicit assumption that was not explicitly tested.

Using a common space approach, however, Eichert, Robinson, et al. (2020) were finally able to dissociate the effects of cortical expansion in the temporal lobe from the effect of tract extension. They demonstrated that the human AF extends further into the temporal lobe than cortical expansion alone would predict suggesting that brain connectivity underwent additional evolutionary modifications. The species differences in the temporal lobe are more pronounced than those in the frontal lobe, despite the frontal expansion during human evolution (Hill et al. 2010; Thiebaut de Schotten et al. 2012). Branching into new territory, the extended AF endowed humans with unique language function, though the precise functional advantage of this extension remains a matter of debate (Balezeau et al. 2020). Network analysis suggested that via the extended AF, the human brain can dedicate more resources to integrate temporal multimodal areas (Ardesch et al. 2019). The human adaptations to the AF, however, are not limited to frontotemporal connectivity. Sierpowska et al. (2022) showed that the human AF connections between the temporal and the parietal lobe are also more complex compared to those in the chimpanzee.

The temporal lobe extension of the AF can now also be studied in a broader context of general remodelling of the brain's long-range connections. Numerous studies have confirmed a left-lateralization of the human AF, which is less prominent or absent in

9780192886491_Book.indb 584 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:04

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

non-human primates (Eichert et al. 2019). On the other hand, some branches of the superior longitudinal fasciculus have been found to be right-lateralized in human. The striking asymmetry of the human dorsal pathways is thought to support a division of labour across the hemispheres, which allows for more specialized processing, critical in the context of language. Evolution of these dorsal language pathways happened in concert with further adaptive changes to the temporal lobe white matter. The ventral language pathway comprises several longitudinal white-matter streams, which can be reconstructed in a homologous manner across species using neuroimaging. This approach revealed that some pathways, such as the course of the inferior fronto-occipital fasiculus (Mars et al. 2016) and the uncinate fasciculus (Folloni et al. 2019), are more similar between human and macaque than previously expected, whilst others, such as the inferior longitudinal pathway (ILF), show distinctly human specializations (Roumazeilles et al. 2020). The ILF, which connects occipital and inferior temporal areas, has a more complex structure in humans, and to a lesser extent in other great apes, compared to macaques, forming several subdivisions.

In addition to the connectivity changes in the temporal lobe, there were major neocortical expansions of the lateral and inferior temporal lobe (Hill et al. 2010; Bryant and Preuss 2018), which affected the sulcal architecture of the lobe. Whilst humans and chimpanzees possess two longitudinal temporal sulci (superior temporal sulcus (STS) and inferior temporal sulcus (ITS)), the ITS is not present in macaques. The human middle temporal gyrus (MTG) is well established as a 'semantic hub' and has been widely implicated in linguistic processing (Dronkers et al. 2004). Whilst functional studies on semantic processing in macaques are rare (Froesel et al. 2022), the existing evidence on multimodal processing suggests that a precursor of the human semantic MTG hub is present inside the macaque STS (Petrides and Iversen 1978). This macaque precursor is thought to have expanded in higher apes and as a consequence the MTG was 'pushed' out forming an additional gyrus (Petrides and Pandya 1988). In line with this hypothesis routed in morphology, the middle temporal areas are those which have a maximal connectional dissimilarity between macaques and humans (Mars, Sotiropoulos, et al. 2018) suggesting that connectivity remodelling happened in concert with the cortical expansions supporting the semantic system.

Whilst the human temporal lobe volume itself is not relatively bigger than that of the macaque, the amount of temporal cortex relative to its primary thalamic inputs is drastically increased in the human when compared to several other primates (Braunsdorf et al. 2021). This indicates that the human temporal cortex devotes more resources to local processing of incoming signals. Taken together with connectional and morphological species differences this points to the temporal lobe as a hot spot for structural adaptations to support higher-order language processes.

Another well-established hotspot for studies on language evolution is the frontal lobe. The cortical expansion of the frontal lobe is a hallmark of evolutionary brain theories (Passingham 2008; Passingham and Wise 2012). Similarly, the role of left ventral lateral frontal lobe, often referred to as Broca's area, as a critical hub for language function is well recognized (Hagoort 2005). A large body of literature has examined the

9780192886491_Book.indb 585 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:05:04

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

anatomical basis supporting the function of Broca's area in humans. Language function in the frontal lobe is typically left-lateralized, but whether underlying structural asymmetry supports this functional lateralization is still a matter of debate (Sprung-Much et al. 2022). The evolution of hemispheric differences in brain structure was the focus of a recent genetic investigation combining genetic and neuroimaging data from the UK Biobank with that from Neanderthals and Denisovans (Alagöz et al. 2022). The asymmetry of brain surface area in language-relevant areas of the left hemisphere was shown to be genetically enriched in modern humans and strongly heritable. Furthermore, a link to lateralization of the uncinate fasciculus was found, but note that the AF was not part of the investigation. Comparable data from non-human primates was not included, but the results can well be placed in context with findings from other data modalities.

Histological investigations demonstrated a comparable blueprint of frontal lobe cortical areas between humans and macaques (Petrides 1999; Petrides et al. 2005). These findings were replicated in a comparative study of ventral lateral cortex using neuroimaging (Neubert et al. 2014, 2015). Profound species differences, however, were found in the functional coupling with posterior auditory association areas. Further studies showed that the coupling of frontal and parietal lobes follow a similar core-shell organization in human and macaques (Margulies and Petrides 2013; Vijayakumar et al. 2019), though species differences in fronto-parietal connectivity have been suggested between humans and macaques (Bryant et al. 2018). These findings strengthen the hypothesis that the function of Broca's area in language is based on a unique connectivity profile in the human, rather than on local changes to brain structure. This hypothesis is in line with a general concept of the brain's architecture put forward by Passingham et al. (2002), namely that each brain area's function is constrained and determined by its unique inputs and outputs.

Studies on the neurobiology of vocalizations and vocal learning offered another comparative perspective on language evolution, which complements and overlaps with some of the findings discussed above. Laryngeal motor control is one aspect of vocal behaviour, which is a long standing focus of evolutionary speech research (Fitch 2010; Ackermann et al. 2014; Simonyan 2014; Belyk and Brown 2017; Eichert, Papp, et al. 2020). Non-human primates have one laryngeal motor representation in the premotor cortex (Hast et al. 1974; Simonyan and Jürgens 2002), and humans have two distinct laryngeal representations. This observation gave rise to the human larynx motor cortex 'duplication' theory (Belyk and Brown 2017; Belyk et al. 2021). As mentioned above, Eichert, Robinson, et al. (2020) found that the ventral laryngeal representation is likely to be homologous to the non-human primate one in the premotor cortex, whilst the dorsal one is located in the primary motor cortex.

The larynx motor cortex duplication theory further predicts that the duplication of the larynx motor cortex is repeated throughout the broader motor network (Belyk et al. 2021), but experimental validation is missing as of yet. The concept of two parallel, or duplicated, vocal motor control pathways has also been described for the avian brain (Chakraborty and Jarvis 2015). In line with this theory, genetic expression profiles in the human dorsal and ventral larynx area have similarities with those in analogous motor nuclei in the avian brain (Pfenning et al. 2014). It has therefore been argued that the

9780192886491_Book.indb 586 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:04

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

evolutionary principle of brain pathway duplication has evolved analogously in vocal learners, such as in songbirds and humans, but not in non-human primates, who have limited vocal learning capacities (Lameira 2017; Jarvis 2019).

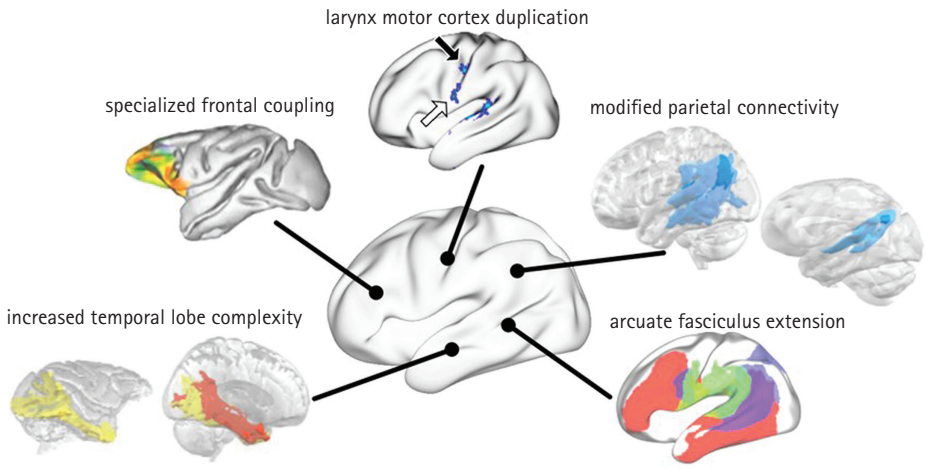


FIGURE 31.2 Summary of human brain adaptations supporting speech and language.

Sources: Larynx motor cortex duplication: adapted from Eichert, Robinson, et al. (2020), under a CC BY 4.0 license; Specialized frontal coupling: adapted from Neubert et al. (2015), with permission; Modified parietal connectivity: adapted from Sierpowska et al. (2022), under a CC BY-NC-ND license, with permission; Increased temporal lobe complexity: adapted from Roumazeilles et al. (2020), under a CC BY 4.0 license; Arcuate fasciculus extension: adapted from Eichert, Robinson, et al. (2020), under a CC BY 4.0 license.

31.5 Conclusion

Studying the human brain adaptations supporting language requires a comparative approach across species. The field recently leaped forward due two major developments: First, multimodal whole-brain data has become available for a large range of species, owing to advances in neuroimaging. Second, new analysis frameworks have evolved that leverage the full potential of the new data sources. Specifically the 'common space approach' provides us with a framework to formulate new testable hypotheses and to study them across the phylogenetic tree. Due to these developments, we have been able to form a more nuanced understanding of language evolution in the brain and were able to revisit and update long-established principles.

Acknowledgements

NE is supported by a Sir Henry Wellcome Postdoctoral Fellowship from the Wellcome Trust [222799/Z/21/Z]. The work of R.B.M. is supported by the Biotechnology and Biological Sciences Research Council (BBSRC) UK [BB/N019814/1], the EPA Cephalosporin Fund,

9780192886491_Book.indb 587 /12_first_proofs/first_pr/xml_for_typesetting 27-Mar-25 23:05:04

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

and a Gravitation grant from the Netherlands Organization for Scientific Research (NWO) [024.001.006]. The Wellcome Centre for Integrative Neuroimaging is supported by core funding from the Wellcome Trust [201319/Z/16/Z and 203139/A/16/Z].

References

Ackermann, H., Hage, S., and Ziegler, W. (2014). 'Brain mechanisms of acoustic communication in humans and nonhuman primates: An evolutionary perspective', *Behavioral and Brain Sciences* 37: 529–604.

Alagöz, G., Molz, B., Eising, E., et al. (2022). 'Using neuroimaging genomics to investigate the evolution of human brain structure', *Proceedings of the National Academy of Sciences USA* 119: e2200638119.

Amunts, K., Lenzen, M., Friederici, A. D., et al. (2010). 'Broca's region: Novel organizational principles and Multiple Receptor Mapping', *PLoS Biology* 8: e1000489.

Amunts, K., Lepage, C., Borgeat, L., et al. (2013). 'BigBrain: An ultrahigh-resolution 3D human brain model', *Science* (1979) 340: 1472–1475.

Ardesch, D. J., Scholtens, L. H., Li, L., et al. (2019). 'Evolutionary expansion of connectivity between multimodal association areas in the human brain compared with chimpanzees', *Proceedings of the National Academy of Sciences USA* 116: 7101–7106.

Assaf, Y., Bouznach, A., Zomet, O., et al. (2020). 'Conservation of brain connectivity and wiring across the mammalian class', *Nature Neuroscience* 23: 805–808.

Autio, J. A., Glasser, M. F., Ose, T., et al. (2020). 'Towards HCP-Style macaque connectomes: 24-Channel 3T multi-array coil, MRI sequences and preprocessing', *NeuroImage* 215: 116800.

Balezeau, F., Wilson, B., Gallardo, G., et al. (2020). 'Primate auditory prototype in the evolution of the arcuate fasciculus', *Nature Neuroscience* 23: 611–614.

Beauchamp, A., Yee, Y., Darwin, B. C., et al. (2022). 'Whole-brain comparison of rodent and human brains using spatial transcriptomics', *Elife* 11.

Belyk, M. and Brown, S. (2017). 'The origins of the vocal brain in humans', *Neuroscience and Biobehavioral Reviews* 77: 177–193.

Belyk, M., Eichert, N., and McGettigan, C. (2021). 'A dual larynx motor networks hypothesis', *Philosophical Transactions of the Royal Society B: Biological Sciences*.

Braunsdorf, M., Blazquez Freches, G., Roumazeilles, L., et al. (2021). 'Does the temporal cortex make us human? A review of structural and functional diversity of the primate temporal lobe', *Neuroscience and Biobehavioural Reviews* 131: 400–410.

Brodmann, K. (1909). *Vergleichende Lokalisationslehre der Grosshirnrinde*. Barth: Verlag von Jann Ambrosius Barth.

Bruner, E. (2019). 'Human paleoneurology: Shaping cortical evolution in fossil hominids', *Journal of Comparative Neurology* 527(10): 1753–1765.

Bryant, K. L., Li, L., and Mars, R. B. (2018). 'A white matter atlas of the chimpanzee brain and quantitative comparison with human and macaque', *Society for Neuroscience Conference San Diego, CA, USA*.

Bryant, K. L. and Preuss, T. M. (2018). 'A comparative perspective on the human temporal lobe', *Digital Endocasts* 239–258.

Bryant, K. L., Ardesch, D. J., Roumazeilles, L., et al. (2021). 'Diffusion MRI data, sulcal anatomy, and tractography for eight species from the Primate Brain Bank', *Brain Structure Function* 226: 2497–2509.

9780192886491_Book.indb 588 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:04

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Campbell, J. S. W. and Pike, G. B. (2014). 'Potential and limitations of diffusion MRI tractography for the study of language', *Brain and Language* 131: 65–73.

Chakraborty, M. and Jarvis, E. D. (2015). 'Brain evolution by brain pathway duplication', *Philosophical Transactions of the Royal Society B: Biological Sciences* 370: 20150056.

Dronkers, N. F., Wilkins, D. P., van Valin, R. D., et al. (2004). 'Lesion analysis of the brain areas involved in language comprehension', *Cognition* 92: 145–177.

Dunbar, R. (2003). 'Evolution of the social brain', *Science* 302(5648): 1160–1161.

Eichert, N., Verhagen, L., Folloni, D., et al. (2019). 'What is special about the human arcuate fasciculus? Lateralization, projections, and expansion', *Cortex* 118.

Eichert, N., Papp, D., Mars, R. B., and Watkins, K. E. (2020). 'Mapping human laryngeal motor cortex during vocalization', *Cerebral Cortex* 30: 6254–6269.

Eichert, N., Robinson, E. C., Bryant, K. L., et al. (2020). 'Cross-species cortical alignment identifies different types of anatomical reorganization in the primate temporal lobe', *Elife* 9.

Eichert, N., DeKraker, J., Howard, A. F. D., et al. (2024). 'Hippocampal connectivity patterns echo macroscale cortical evolution in the primate brain', *Nature Communications* 15: 5963.

Elston, G. N. (2007). 'Specialization of the neocortical pyramidal cell during primate evolution', in T. M. Preuss (ed.), *Evolution of Nervous Systems*, vol. 4, 191–242. Amsterdam: Elsevier.

Fitch, W. T. (2010). *The Evolution of Language*. Cambridge: Cambridge University Press.

Foley, N. M., Springer, M. S., and Teeling, E. C. (2016). 'Mammal madness: Is the mammal tree of life not yet resolved?', *Philosophical Transactions of the Royal Society B: Biological Sciences* 371(1699): 20150140.

Folloni, D., Sallet, J., Khrapitchev, A. A., et al. (2019). 'Dichotomous organization of amygdala/temporal-prefrontal bundles in both humans and monkeys', *Elife* 8.

Friederici, A. D. (2017). *Language in Our Brain: The Origins of a Uniquely Human Capacity*. Cambridge, MA: MIT Press.

Friedrich, P., Forkel, S. J., Amiez, C., et al. (2020). 'Imaging the primate brain evolution: The next frontier?', *NeuroImage* 13: 16–17.

Froesel, M., Gacoin, M., Clavagnier, S., et al. (2022). 'Socially meaningful visual context either enhances or inhibits vocalisation processing in the macaque brain', *Nature Communications* 13: 1–17.

Froudist-Walsh, S., Xu, T., Niu, M., et al. (2023). 'Gradients of receptor expression in the macaque cortex', *Nature Neuroscience* 26: 1281–1294.

Glasser, M. F., Preuss, T. M., Nair, G., et al. (2012). 'Improved cortical myelin maps in humans, chimpanzees, and macaques allow identification of putative areal homologies', *Society for Neuroscience* 97312.

Hagoort, P. (2005). 'On Broca, brain, and binding: A new framework', *Trends in Cognitive Science* 9: 416–423.

Hast, M. H., Fischer, J. M., Wetzel, A. B., and Thompson, V. E. (1974). 'Cortical motor representation of the laryngeal muscles in Macaca mulatta', *Brain Research* 73: 229–240.

Hayashi, T., Hou, Y., Glasser, M. F., et al. (2021). 'The nonhuman primate neuroimaging and neuroanatomy project', *Neuroimage* 229: 117726.

Healy, S. D. (2021). 'Adaptation and the brain', in Susan D. Healy (ed.), *Adaptation and the Brain*, 1–162. New York: Oxford University Press.

Heuer, K., Gulban, O. F., Bazin, P.-L. L., et al. (2019). 'Evolution of neocortical folding: A phylogenetic comparative analysis of MRI from 34 primate species', *Cortex* 118: 275–291.

9780192886491_Book.indb 589 /12_first_proofs/first_pr[illegible]/xml_for_typesetting 27-Mar-25 23:05:04

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Hill, J., Inder, T., Neil, J., et al. (2010). 'Similar patterns of cortical expansion during human development and evolution', *Proceedings of the National Academy of Sciences USA* 107: 13135–13140.

Howard, A. F. D., Huszar, I. N., Smart, A., et al. (2023). 'An open resource combining multi-contrast MRI and microscopy in the macaque brain', *Nature Communications* 14: 4320.

Jarvis, E. D. (2004). 'Learned birdsong and the neurobiology of human language', *Annals of the New York Academy of Sciences* 1016: 749–777.

Jarvis, E. D. (2019). 'Evolution of vocal learning and spoken language', *Science* (*1979*) 366: 50–54.

Jones, A. R., Overly, C. C., and Sunkin, S. M. (2009). 'The Allen brain atlas: 5 years and beyond', *Nature Reviews Neuroscience* 10(11): 821–828.

Kaas, J. H. and Van Eden, C. G. (2011). 'Brain banks provide a valuable resource for comparative studies', *Brain Behavior Evolution* 77: 65–66.

Keil, J. M., Qalieh, A., and Kwan, K. Y. (2018). 'Brain transcriptome databases: A user's guide', *Journal of Neuroscience* 38: 2399–2412.

Lameira, A. R. (2017). 'Bidding evidence for primate vocal learning and the cultural substrates for speech evolution', *Neuroscience and Biobehavioral Reviews* 83: 429–439.

Lein, E. S., Hawrylycz, M. J., Ao, N., et al. (2006). 'Genome-wide atlas of gene expression in the adult mouse brain', *Nature* 445(7124): 168–176.

Lerch, J. P., van der Kouwe, A. J. W., Raznahan, A., et al. (2017). 'Studying neuroanatomy using MRI', *Nature Neuroscience* 20(3) (2017): 314–326.

Margulies, D. S. and Petrides, M. (2013). 'Distinct parietal and temporal connectivity profiles of ventrolateral frontal areas involved in language production', *Journal of Neuroscience* 33: 16846–16852.

Margulies, D. S., Vincent, J. L., Kelly, C., et al. (2009). 'Precuneus shares intrinsic functional architecture in humans and monkeys', *Proceedings of the National Academy of Sciences USA* 106: 20069–20074.

Mars, R. B., Jbabdi, S., Sallet, J., et al. (2011). 'Diffusion-weighted imaging tractography-based parcellation of the human parietal cortex and comparison with human and macaque resting-state functional connectivity', *Journal of Neuroscience* 31: 4087–4100.

Mars, R. B., Neubert, F. X., Verhagen, L., et al. (2014). 'Primate comparative neuroscience using magnetic resonance imaging: Promises and challenges', *Frontiers in Neuroscience* 8: 1–11.

Mars, R. B., Foxley, S., Verhagen, L., et al. (2016). 'The extreme capsule fiber complex in humans and macaque monkeys: A comparative diffusion MRI tractography study', *Brain Structure and Function* 221: 4059–4071.

Mars, R. B., Eichert, N., Jbabdi, S., et al. (2018). 'Connectivity and the search for specializations in the language-capable brain', *Current Opinion Behavioral Sciences* 21: 19–26.

Mars, R. B., Passingham, R. E., and Jbabdi, S. (2018). 'Connectivity fingerprints: From areal descriptions to abstract spaces', *Trends in Cognitive Science* 22: 1026–1037.

Mars, R. B., Sotiropoulos, S. N., Passingham, R. E., et al. (2018). 'Whole brain comparative anatomy using connectivity blueprints', *Elife* 7.

Mars, R. B., Jbabdi, S., and Rushworth, M. F. S. (2021). 'A common space approach to comparative neuroscience', *Annual Review of Neuroscience* 44: 69–86.

Milham, M. P., Ai, L., Koo, B., et al. (2018). 'An open resource for non-human primate imaging', *Neuron* 100: 61–74.e2.

Neubert, F.-X., Mars, R. B., Thomas, A. G., et al. (2014). 'Comparison of human ventral frontal cortex areas for cognitive control and language with areas in monkey frontal cortex', *Neuron* 81: 700–713.

9780192886491_Book.indb 590 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:04

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Neubert, F.-X., Mars, R. B., Sallet, J., and Rushworth, M. F. S. (2015). 'Connectivity reveals relationship of brain areas for reward-guided learning and decision making in human and monkey frontal cortex', *Proceedings of the National Academy of Sciences USA* 112: E2695–E2704.

Paquola, C., Royer, J., Lewis, L. B., et al. (2021). 'The BigBrainWarp toolbox for integration of BigBrain 3D histology with multimodal neuroimaging', *Elife* 10.

Passingham, R. (2008). *What's Special About the Human Brain*. Oxford: Oxford University Press.

Passingham, R. E. and Wise, S. P. (2012). *The Neurobiology of the Prefrontal Cortex*. Oxford: Oxford University Press.

Passingham, R. E., Stephan, K. E., and Kötter, R. (2002). 'The anatomical basis of functional localization in the cortex', *Nature Reviews Neuroscience* 3: 606–616.

Petrides, M. (1999). 'Dorsolateral prefrontal cortex: Comparative cytoarchitectonic analysis in the human and the macaque brain and corticocortical connection patterns', *European Journal of Neuroscience* 11: 1011–1036.

Petrides, M. and Iversen, S. D. (1978). 'The effect of selective anterior and posterior association cortex lesions in the monkey on performance of a visual-auditory compound discrimination test', *Neuropsychologia* 16: 527–537.

Petrides, M. and Pandya, D. N. (1988). 'Association fiber pathways to the frontal cortex from the superior temporal region in the rhesus monkey', *Journal of Comparative Neurology* 273: 52–66.

Petrides, M., Cadoret, G., and Mackey, S. (2005). 'Orofacial somatomotor responses in the macaque monkey homologue of Broca's area', *Nature* 435: 1235–1238.

Pfenning, A. R., Hara, E., Whitney, O., et al. (2014). 'Convergent transcriptional specializations in the brains of humans and song-learning birds', *Science* (*1979*) 346: 1256846–1256846.

Reveley, C., Seth, A. K., Pierpaoli, C., et al. (2015). 'Superficial white matter fiber systems impede detection of long-range cortical connections in diffusion MR tractography', *Proceedings of the National Academy of Sciences USA* 112: E2820–E2828.

Rilling, J. K. and Insel, T. R. (1998). 'Evolution of the cerebellum in primates: Differences in relative volume among monkeys, apes and humans', *Brain Behaviour and Evolution* 52: 308–314.

Rilling, J. K. and Seligman, R. A. (2002). 'A quantitative morphometric comparative analysis of the primate temporal lobe', *Journal of Human Evolution* 42: 505–533.

Rilling, J. K., Glasser, M. F., Preuss, T. M., et al. (2008). 'The evolution of the arcuate fasciculus revealed with comparative DTI', *Nature Neuroscience* 11: 426–428.

Rilling, J. K., Glasser, M. F., Jbabdi, S., et al. (2012). 'Continuity, divergence, and the evolution of brain language pathways', *Frontiers in Evolutionary Neuroscience* 3: 1–6.

Roumazeilles, L., Eichert, N., Bryant, K. L., et al. (2020). 'Longitudinal connections and the organization of the temporal cortex in macaques, great apes, and humans', *PLoS Biology* 18: e3000810.

Sallet, J., Mars, R. B., Noonan, M. P., et al. (2011). 'Social network size affects neural circuits in macaques', *Science* (*1979*) 334: 697–700.

Sierpowska, J., Bryant, K. L., Janssen, N., et al. (2022). 'Comparing human and chimpanzee temporal lobe neuroanatomy reveals modifications to human language hubs beyond the frontotemporal arcuate fasciculus', *Proceedings of the National Academy of Sciences* 119: e2118295119.

9780192886491_Book.indb 591 /12_first_proofs/first_pr/xml_for_typesetting 27-Mar-25 23:05:04

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Simonyan, K. (2014). 'The laryngeal motor cortex: Its organization and connectivity', *Current Opinions in Neurobiology* 28: 15–21.

Simonyan, K. and Jürgens, U. (2002). 'Cortico-cortical projections of the motorcortical larynx area in the rhesus monkey', *Brain Research* 949: 23–31.

Sprung-Much, T., Eichert, N., Nolan, E., and Petrides, M. (2022). 'Broca's area and the search for anatomical asymmetry: Commentary and perspectives', *Brain Structure and Function* 227: 441–449.

Stephan, H., Frahm, H., and Baron, G. (1981). 'New and revised data on volumes of brain structures in insectivores and primates', *Folia Primatologica* 35: 1–29.

Tendler, B. C., Hanayik, T., Ansorge, O., et al. (2022). 'The Digital Brain Bank, an open access platform for post-mortem datasets', *Elife* 11.

Thiebaut de Schotten, M., Dell'Acqua, F., Valabregue, R., and Catani, M. (2012). 'Monkey to human comparative anatomy of the frontal lobe association tracts', *Cortex* 48: 82–96.

Thiebaut de Schotten, M., Croxson, P. L., and Mars, R. B. (2019). 'Large-scale comparative neuroimaging: Where are we and what do we need?', *Cortex* 118: 188–202.

van Essen, D. C. and Dierker, D. L. (2007). 'Surface-based and probabilistic atlases of primate cerebral cortex', *Neuron* 56: 209–225.

Vijayakumar, S., Sallet, J., Verhagen, L., et al. (2019). 'Mapping multiple principles of parietal–frontal cortical organization using functional connectivity', *Brain Structure and Function* 224: 681–697.

Vincent, J. L., Patel, G. H., Fox, M. D., et al. (2007). 'Intrinsic functional architecture in the anaesthetized monkey brain', *Nature* 447: 83–86.

Weiskopf, N., Suckling, J., Williams, G., et al. (2013). 'Quantitative multi-parameter mapping of R1, PD*, MT, and R2* at 3T: A multi-center validation', *Frontiers in Neuroscience* 7: 95.

Xu, T., Nenning, K. H., Schwartz, E., et al. (2020). 'Cross-species functional alignment reveals evolutionary hierarchy within the connectome', *Neuroimage* 223: 117346.

Zilles, K. and Amunts, K. (2010). 'Centenary of Brodmann's map conception and fate', *Nature Reviews Neuroscience* 11(2): 139–145.

9780192886491_Book.indb 592 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:04

CHAPTER 32

HOW TO STUDY ANIMAL SYNTAX

TOSHITAKA N. SUZUKI

32.1 Syntax in animals?

Syntax is defined as a set of rules that determines whether sequences of words are well formed (Chomsky 2001; Hurford 2011). For example, 'birds catch caterpillars' is a well-formed sentence, while 'catch birds caterpillars' is not. When discriminating between these sequences, both senders and receivers use a rule (e.g. Subject → Verb → Object) to produce and understand the relationship between the three words. Syntax not only determines the relationship between multiple words, but also enhances the expressive power of language. For example, from the same three words, we can create a sentence 'caterpillars catch birds.' Although this describes an unrealistic situation, its meaning can be comprehended by applying the same rule. From a semantic perspective, syntactically combined word sequences follow the principle of compositionality (also known as Frege's principle), by which the meaning of a complex expression depends on the meanings of its constituent parts and the rules by which they are combined (Partee et al. 1990; Pelletier 1994). With some exceptions, such as idioms (e.g. 'to kick the bucket' means 'to die'), the meaning of phrases or sentences is, in principle, derived from the meanings of component words.

Zoologists have long used the term 'syntax' to describe combinations of vocal elements in non-human species, such as passerines, whales, and gibbons (Podos et al. 2004; Clarke et al. 2006; Catchpole and Slater 2008; Allen et al. 2019). These animals produce sequences of multiple sound elements (so-called songs) while advertising their territories or attracting mates. However, these phenomena are not included in the standard definition of syntax because the vocal elements within these sequences are generally meaningless, and the combined sequences have the simple function of advertising features associated with the signaller, such as identity or quality (Catchpole and Slater 2008). In other words, even if a sequence is structurally complex and follows a specific ordering rule, it does not provide a parallel with human syntax in the perspective of its semantic

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

construction. As the term 'syntax' is used in linguistics to describe the combinations of meaning-bearing units, it is also necessary to use syntax to describe the sequences of meaning-bearing vocal units in animal studies (Suzuki and Zuberbühler 2019).

Despite the historical assumption that syntax is unique to humans, recent advances in field research have revealed that non-human animals may have evolved similar capabilities (Suzuki et al. 2018, 2019; Zuberbühler 2019; Suzuki 2021). Strong evidence for syntax has been provided by a wild passerine, the Japanese tit (*Parus minor*). This bird species has evolved a variety of call types and uses them for social interactions, such as recruiting flock members or warning offspring about predators (Suzuki 2014; Suzuki et al. 2016). A series of field experiments have shown that Japanese tits have evolved syntactic communication; they (i) combine two meaning-bearing calls into ordered sequences, (ii) extract compositional meanings from call sequences, (iii) recognize call orderings, (iv) recognize a call sequence as a (semantically) bound construction, and (v) use an ordering rule to decode novel call sequences (Suzuki 2014; Suzuki et al. 2016, 2017; Suzuki and Matsumoto 2022). Using these studies on Japanese tits as examples, this chapter illustrates general paradigms for studying animal syntax.

32.2 Methodology

32.2.1 Field observations and call recording

The initial step in exploring the capacity for animal syntax is to investigate the vocal production of the study species. It is therefore necessary to quantify how animal species use different call types in different contexts. One such method is field observation. For example, many birds produce different call types in different behavioural contexts, such as foraging and flight, which could be recorded in a natural setting (Freeberg 2008; Freeberg and Mahurin 2013). However, some call types may be difficult to record in certain situations. For example, many birds and mammals give alarm calls when they encounter predators, but it is not easy to record all the details of predator encounters from the onset in a natural setting. Additionally, the acoustic structure of call types may be influenced by the type, distance, behaviour, and speed of the predator (Seyfarth et al. 2010; Suzuki 2016), making it difficult to assess the factors that influence alarm call structure. Field experiments are useful in such cases; for example, vocal responses can be evoked by experimentally exposing animals to predator models (Seyfarth et al. 2010; Suzuki 2016). Similarly, some birds and mammals produce vocalizations when finding a food source; such vocalizations (so-called 'food-associated calls') can be experimentally elicited with the provisioning of food sources to wild animals (Clay et al. 2012).

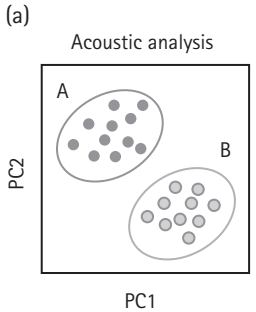
Once the vocal data are collected, it is necessary to analyse the proportion or number of each call type produced in each context. Classification of individual call types can be conducted by multivariate analyses (e.g. principal component or cluster analyses) or deep machine learning procedures (Ghani et al. 2023). Then, by examining the distribution of intervals between individual calls (e.g. bout length analyses), call sequences could be determined.

9780192886491_Book.indb 594 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:04

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

When sequences of call types are detected, it can be analysed whether these sequences are responsive to specific situations or represent randomly produced call sequences (see Chapter 28). In the case of an animal species that has evolved syntax, individual call types should have distinct meanings, thus enabling their use in different contexts. Consequently, it is expected that different calls (A and B) and their combined sequence (AB) are produced differently across the situations and/or that different call sequences (AB and BA) are used in different situations. It is also worth exploring the existence of an ordering rule to combine call types. For example, AB sequences may be used more frequently than BA sequences in a particular situation, suggesting that animals have evolved an ordering rule (A → B) (Figure 32.1). However, there might be a case where AB and BA are produced at the same frequency but with different meanings. For instance, consider a situation where 'green light' and 'light green' appear with the same frequency in a language. In such cases, there exists a rule of adjective–noun, but there appears to be no rule to determine the sound ordering.

A combination of field observations and predator model presentations were used to determine how Japanese tits produce separate calls and combined call sequences (Suzuki 2014; Suzuki et al. 2016; Suzuki and Matsumoto 2022). These birds produce alert calls when warning conspecifics about danger, such as the presence of predators, whereas they produce acoustically distinct recruitment calls when attracting conspecifics to non-dangerous situations, such as food locations or nests. These call types are often combined into ordered sequences (alert-recruitment call sequences) when gathering other individuals to approach and harass (i.e. mob) a stationary predator (Figure 32.2). By analysing the intervals between call types (or note types), it could be determined whether different call types are combined or not, e.g. in the case of Japanese tits different call types are typically combined into sequences with an interval of 50–150 ms. However, they do not combine alert and recruitment calls into recruitment–alert call sequences. Such sound ordering rules suggest that Japanese tits may have evolved a syntactic ordering rule (alert → recruitment) for combining calls (see also Schlenker et al. 2016a, b).



(b)

Call usage and ordering rule

	Context 1	Context 2
A	10%	60%
B	20%	30%
A-B	70%	10%
B-A	0%	0%

FIGURE 32.1 Investigation of call production.

(a) One of the first steps in testing for call sequences is to classify recorded calls into categories. Multivariate analyses (e.g. principal component or cluster analyses) could be a useful way to classify call categories. (b) Next, the exploration can focus on how different call types and their combined sequences are produced. This also facilitates investigations into ordering rules for call combinations.

9780192886491_Book.indb 595 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:05:05

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

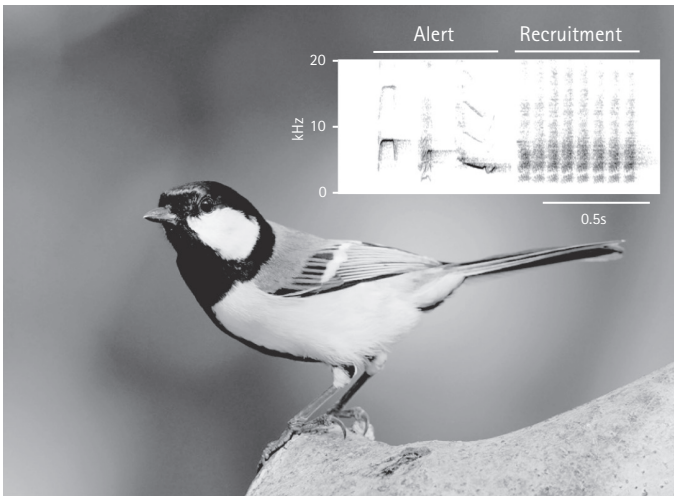


FIGURE 32.2 Japanese tits combine alert and recruitment calls into alert-recruitment call sequences when gathering other flock members to approach and harass stationary predators, such as shrikes.

32.2.2 Playback of solo vs. combined calls

If an animal uses calls A and B individually and in combination as AB sequences, playback experiments should be designed to test whether solo and combined call sequences evoke different reactions in receivers. If receivers detect the compositional meaning of A and B from combined AB sequences, they are expected to respond differently to the A, B, and AB. Moreover, their responses to AB might be related to those of both A and B if the combination follows the principle of compositionality. To ensure that all call stimuli evoke behavioural reactions in receivers, it is necessary to prepare control trials, such as playback of background noise, which should not evoke responses (Figure 32.3a).

As a practical example of these methods, playback experiments were conducted on Japanese tits with alert calls, recruitment calls, alert-recruitment call sequences, and background noise control (Suzuki et al. 2016). Japanese tit receivers display different behaviours when hearing alert calls (moving their head horizontally as if scanning for danger) and recruitment calls (approaching the sound source). In response to alert–recruitment call sequences, they progressively approach the sound source while continuously scanning the horizon, suggesting that they detect compositional information (i.e. 'alert' + 'approach') (Figure 32.4).

When designing playback experiments, certain issues should be considered during stimulus preparation. First, pseudo-replication must be avoided (Kroodsma et al. 2001); if playback treatments are compared in order to test a hypothesis, it is necessary to prepare multiple exemplars for each treatment. If only one playback exemplar is used for each treatment, it cannot be distinguished by whether responses are caused by

9780192886491_Book.indb 596 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:05

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

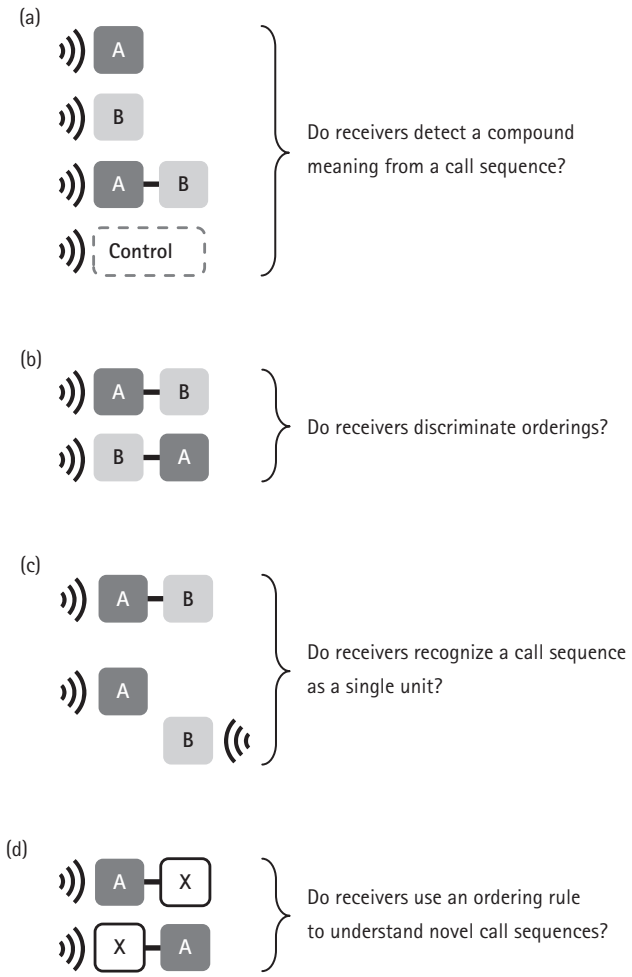


FIGURE 32.3 Schematic representation for playback experiments.

(**a**) Evaluation of receivers' reactions to call stimuli A, B, and AB, with control trials assessing behavioural responses compared to baseline behavioural activities. (**b**) Examination of receivers' responses to naturally ordered (AB) and artificially reversed (BA) call sequences. (**c**) Assessment of receivers' ability to recognize temporally linked calls (AB) originating from the same or different sources. (**d**) Investigation into receivers' use of an ordering rule through the creation of novel call sequences with different orderings (A→X or X→A), unveiling their capacity to decode even novel combinations of known calls.

a particular exemplar or by a particular call type (i.e. a given treatment). The best solution for avoiding pseudo-replication is to prepare the number of playback exemplars equal to the number of trials per treatment. Second, it is necessary to control for the possible influence of the individual identification of call stimuli. In birds and mammals, it may be common for individuals to recognize other individuals based on their acoustic characteristics (Carlson et al. 2020). Therefore, familiarity with source individuals

9780192886491_Book.indb 597 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:06

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

should be controlled when preparing sound stimuli or when making experimental designs. Third, one should consider potential geographic variation in call structure; in some birds and mammals, acoustic characteristics of calls vary across geographic regions (Janik and Knörnschild 2021; Williams and Lachlan 2021). Therefore, it is preferable to use call stimuli recorded in the same region as the experimental site. Finally, it is necessary to ensure data independency. For example, if each individual was tested ten times for each treatment, such individual replications should be considered in the statistical analysis. In order to avoid habituation effects for the same stimuli, it would be better to increase the number of subject animals and reduce the number of playback trials per each individual.

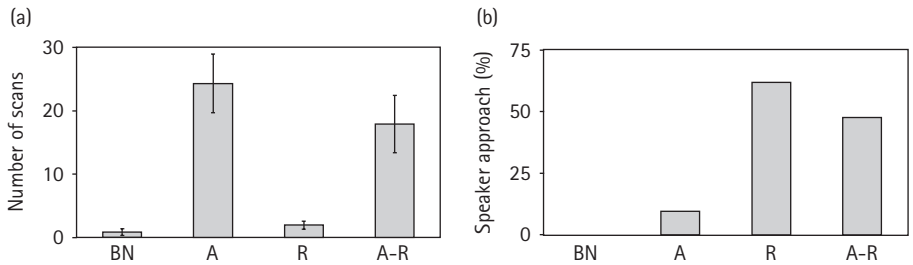


FIGURE 32.4 Japanese tits move their heads horizontally in response to playback of alert calls, approach the speaker in response to recruitment calls, and combine these two behaviours in response to alert–recruitment call sequences.

(**a**) Number of horizontal scans during playbacks, (**b**) Proportion of birds that approach within 2 m of the speaker. Abbreviations: BN: background noise (control); A: alert calls; R: recruitment calls; A-R: alert–recruitment call sequences.

Source: Redrawn from Suzuki et al. 2016.

32.2.3 Testing rules

To ascertain whether animal call sequences are syntactically combined, it is essential to test whether rules impact the meaning of the call sequences. An apparent rule is ordering: if an animal combines A and B calls into AB but not BA sequences, playback experiments could be employed to assess whether receivers can discriminate between these two sequences (Figure 32.3b). Japanese tits combine alerts with recruitment calls in an ordered sequence (Suzuki 2014). Therefore, a playback experiment was designed to test whether receivers discriminate between naturally ordered call sequences (alert → recruitment) and artificially reversed call sequences (recruitment → alert) (Suzuki et al. 2016). The results show that Japanese tit receivers progressively approach the sound source while scanning the horizon in response to alert–recruitment call sequences; however, they reduce their responses when hearing recruitment–alert call sequences. This suggests that tits have a capacity to recognize whether the component calls are temporally linked into a specific ordering (Figure 32.5).

9780192886491_Book.indb 598 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:06

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

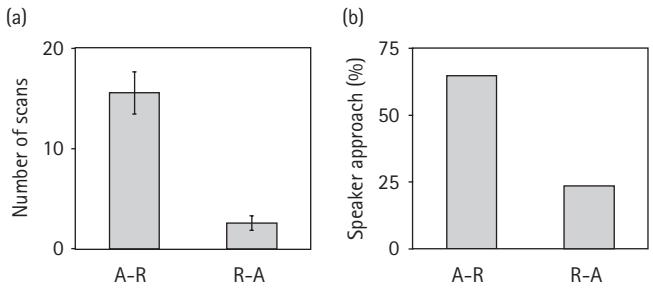


FIGURE 32.5 Japanese tits discriminate between naturally ordered alert–recruitment call sequences and artificially reversed recruitment–alert call sequences. (**a**) Number of horizontal scans during playbacks, (**b**) Proportion of birds that approach within 2 m of the speaker. Abbreviations: A-R: alert–recruitment call sequences; R-A: recruitment–alert call sequences.

Source: Redrawn from Suzuki et al. 2016.

32.2.4 Recognition of a combined unit

Although signallers combine calls into a sequence and receivers exhibit compositional responses to naturally ordered call sequences, there is an alternative explanation that does not depend on syntax. Receivers may perceive a call sequence as two individual calls arbitrarily produced in close time proximity (i.e. temporally linked, A and B) and not as a single structure comprised of two call types (i.e. combined {AB}). If receivers have the capacity to recognize a two-call sequence as a single structure produced by a single individual, they should perceive whether component calls originate from the same spatial location in a temporally linked manner. This can be tested by broadcasting temporally linked two calls from one or two sources (Figure 32.3c).

This paradigm was used to test whether Japanese tit receivers recognize an alert–recruitment call sequence as a single structure comprised of two call types (Suzuki and Matsumoto 2022). The results show that receivers mob a predator specimen when hearing alert–recruitment call sequences from a single speaker. However, they do not exhibit predator mobbing when hearing the same two calls broadcast from different sources (Figure 32.6). Therefore, Japanese tits process the alert–recruitment call sequence as a single compositional message to prompt others to mob a predator. Of course, there might be cases where individuals detect compositional information when hearing two words from different individuals (e.g. Ann says 'it's raining' and Bill says 'it's windy' and you will be able to process this information as 'I go out with my raincoat rather than my umbrella which won't hold against the wind'). However, if there are cases where animals respond differently when hearing the same two calls broadcast from one or different sources, this could provide evidence that they have the ability to process a sequence of two calls as a single structure produced by a single individual.

9780192886491_Book.indb 599 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:05:06

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Several points should be noted while using this experimental paradigm. If two call types are recorded from a single source individual, it would be unrealistic to broadcast the two calls separately from two speakers. In contrast, if two call types are recorded from two source individuals, it would be unrealistic to broadcast the two calls from a single source. Therefore, researchers should prepare two types of call combinations (two call types from either one or two source individuals) to statistically control the effects of possible individual-based signatures in call structures. In the experiments on Japanese tit call sequences, calls from either one or two unfamiliar birds were used, and there was no influence from the number of source individuals (one vs. two) on mobbing behaviour towards the predator specimen during playbacks (Suzuki and Matsumoto 2022).

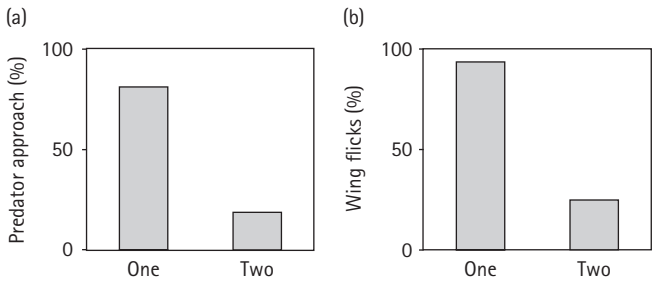


FIGURE 32.6 Japanese tits mob a shrike specimen in response to alert–recruitment call sequences broadcast from a single speaker.

However, they do not mob the shrike specimen when the component calls are broadcast from two different speakers, even though these two calls are broadcast in a temporally linked manner as in the one-speaker treatment. (**a**) Percentage of trials in which at least one Japanese tit within a flock approached within 2 m of the shrike specimen, (**b**) Percentage of trials in which at least one Japanese tit within a flock exhibited wing flicking displays. Abbreviations: One: one-speaker playback; Two: two-speaker playbacks.

Source: Redrawn from Suzuki and Matsumoto 2022.

32.2.5 Recognition of novel call sequences

The expressive power of language depends on the syntax by which individuals produce and understand novel combinations of words (Chomsky 1965). If an animal species has evolved such a flexible use of syntax, they should be able to use their syntactic rules to understand novel combinations of known calls. However, finding novel call combinations in the natural setting is challenging as all call sequences recorded in the wild are not novel to wild animals. Even if the call combination is artificially edited into a novel order that has not been recorded in the wild population, there are no guarantees that such a call sequence has ever been used by the subject animals throughout their lifetimes. Moreover, it is possible that such edited sequences do not follow an ordering rule and, therefore, cannot be recorded naturally.

9780192886491_Book.indb 600 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:07

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

One solution is to mix vocalizations recorded from different animal species, thereby creating novel mixed-species call sequences (Figure 32.3d). For example, Japanese tits form mixed-species flocks with willow tits (*Poecile montanus*) during the non-breeding season (Suzuki 2012). In these flocks, Japanese tits approach in response to the recruitment calls of both conspecifics and willow tits in order to maintain flock cohesion (Suzuki 2012; Suzuki and Kutsukake 2017). However, in a playback experiment where willow tit recruitment calls were artificially shortened to match the length of Japanese tit recruitment calls (while maintaining their natural pitch), Japanese tits did not approach these modified calls (Suzuki et al. 2017). This result suggests that Japanese tit responses are not induced by acoustic similarity between the recruitment calls of willow tits and their own species, but rather because they perceive them as distinct vocalizations with a shared meaning ('approach'). In other words, from the perspective of Japanese tit receivers, willow tit recruitment calls are synonymous with the recruitment calls of their own species. This provides the opportunity to generate novel artificial call sequences composed of conspecific alert calls and heterospecific recruitment calls. Playback experiments showed that Japanese tits approach alertness to both natural and novel (mixed-species) call sequences only when the combinations of call units follow the alert–recruitment ordering, indicating that they use an ordering rule to decode novel call combinations (Figure 32.7; Suzuki et al. 2017). This provides evidence of an ability to generalize syntactic processing capabilities, which is key to linguistic productivity.

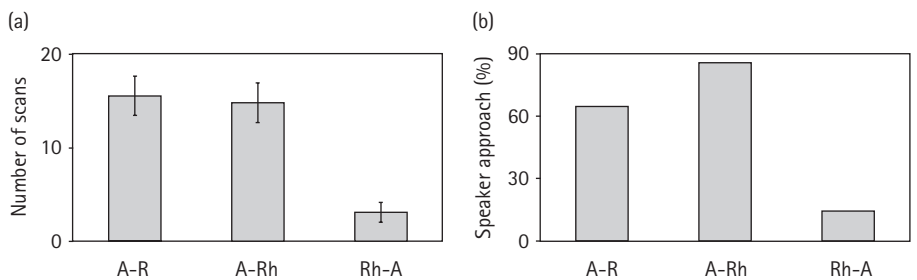


FIGURE 32.7 Japanese tits exhibit similar responses to playbacks of their own alert–recruitment call sequences (A-R) and artificially generated, mixed-species call sequences composed of conspecific alert calls and heterospecific recruitment calls (A-Rh).

However, they reduce their response when mixed-species call sequences do not follow an alert–recruitment ordering rule (Rh-A), suggesting that they can use an ordering rule to decode even novel combinations of calls. (**a**) Number of horizontal scans during playbacks, (**b**) Proportion of birds that approach within 2 m of the speaker. Abbreviations: A-R: alert–recruitment call sequences; A-Rh: sequences of conspecific alert calls and heterospecific recruitment calls; Rh-A: sequences of heterospecific recruitment calls and conspecific alert calls.

Source: Redrawn from Suzuki et al. (2017).

9780192886491_Book.indb 601 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:05:07

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

32.3 Future directions

32.3.1 Application

To date, combinations of meaning-bearing call types have been documented in various species, including non-human primates (e.g. Ouattara et al. 2009; Leroux et al. 2023) and birds (e.g. Engesser et al. 2016; Suzuki et al. 2016). While some studies have examined receivers' responses to separate calls and combined call sequences (e.g. Coye et al. 2015; Engesser et al. 2016; Leroux et al. 2023), there is a lack of testing whether these animals use rules to process call sequences (Schlenker et al. 2023). Therefore, it remains largely unclear whether these species have evolved syntax (i.e. a set of rules) that determines whether their call sequences are well formed.

The methodology used for Japanese tits may help uncover whether these animals have evolved syntax to perceive call sequences. First, it is not challenging to test the receivers' response to natural and reversed call orderings. Second, it is also easy to assess their response to one- vs. two-speaker playbacks of temporally linked two calls. Finally, the creation of novel mixed-species call sequences is likely applicable across a broad spectrum of animals, as an increasing number of studies indicate that many birds and mammals learn to associate heterospecific vocalizations with specific meanings (Magrath et al. 2015). I hope the methodology introduced in this chapter encourages future studies to examine how animals perceive call sequences, which may uncover how commonly syntax is involved in animal communication systems.

32.3.2 Hierarchical structure

Syntactic communication in humans has been hypothesized to depend on a cognitive capability called Merge (Chomsky 2001; Fujita 2014; Rizzi 2016). One hypothesis suggests that Merge enables us to produce and comprehend any kind of word combination, including complex expressions with hierarchical structures (for example, a + dog + barks = {a dog} + barks = {a dog {barks}}) (Berwick and Chomsky 2016). While it is unclear whether non-human animals can combine more than two meaning-bearing calls into a hierarchical structure, this could be explored using the same methodology used for two-call combinations. For example, if the AB sequence is further combined with C to form an ABC sequence, whether the receivers perceive this as a hierarchical structure can be tested by considering AB as a single unit and then applying the same method to the AB + C combination. Although this requires greater research effort than two-call combinations, it is worthwhile because current linguistic theory assumes that Merge (or recursive Merge; Fujita 2014) is a key feature that distinguishes human language from other animal communication systems.

9780192886491_Book.indb 602 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:07

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

32.3.3 Trivial vs. non-trivial compositionality

Recent field studies on various animal species have revealed diversity in the meanings generated by call combinations (Engesser and Townsend 2019; Berthet et al. 2022). In the case of Japanese tits, call sequences are compositional but simply provide a juxtaposed message (i.e. 'alertly' + 'approach' = 'alertly approach'). This seems to be analogous to human cases, such as 'come' + 'talk' = 'come talk'. However, in human language, combinations of words often convey non-additive information (e.g. 'dog' + 'walking' = 'dog walking'). Whether animal call sequences create a non-juxtaposed message is not yet clear; however, there may be cases in which component calls interact to provide such a non-trivial message (Engesser and Townsend 2019). Future detailed studies that apply the methodology presented here will deepen our understanding of the diversity of meaning creation through syntactic communication in non-human animals.

32.3.4 Signaller's cognition

While playback experiments are useful to explore cognitive processes underlying call perception, it is not easy to reveal cognitive mechanisms underlying signal production. For example, even if a signaller combines two calls into an ordered sequence, such call sequences might be produced for anatomical reasons, not by the application of an internal syntactic ordering rule. If it is possible to train an animal to use an ordering rule to produce novel call combinations, it could provide evidence that the animal has evolved a syntax to produce sequences of meaning-bearing call units. However, such training might require operant conditioning techniques and thus would be difficult to apply in wild animals. For wild populations, it may be helpful to examine whether dialects exist in the ordering rules of call sequences. If there are two populations that are not genetically different, but use different ordering rules to combine calls, this would support the hypothesis that individuals learn an ordering rule for producing call sequences (see Girard-Buttoz et al. 2022). Therefore, investigations of learning and ontogeny are important for exploring the cognitive processes of call production, which is a promising avenue for future studies.

32.4 Conclusion

This chapter illustrates the methodology used to study syntax in non-human animals. Although syntax has long been considered unique to humans, growing evidence suggests that birds and mammals combine meaning-bearing units into sequences. Elucidating the cognitive processes used for the production and recognition of animal

9780192886491_Book.indb 603 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:05:07

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

vocal sequences is not easy but promises to provide a better understanding of the diversity and adaptation of syntactic communication, as well as to establish the theory of language evolution.

References

Allen, J., Garland, E. C., Dunlop, R. A., and Noad, M. J. (2019). 'Network analysis reveals underlying syntactic features in a vocally learnt mammalian display, humpback whale song', *Proceedings of the Royal Society B* 286: 20192014.

Berthet, M., Coye, C., Dezecache, G., and Kuhn, J. (2022). 'Animal linguistics: A primer', *Biology Review* 98: 81–98.

Berwick, R. and Chomsky, N. (2016). *Why Only Us? Language and Evolution*. London: MIT Press.

Carlson, N. V., Kelly, E. M., and Couzin, I. (2020). 'Individual vocal recognition across taxa: a review of the literature and a look into the future', *Philosophical Transactions of the Royal Society B* 375: 1803.

Catchpole, C. K. and Slater, P. J. (2008). *Bird Song: Biological Themes and Variations*. Cambridge: Cambridge University Press.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (2001). 'Derivation by phase', in M. Kenstowicz (ed.), *Ken Hale: A Life in Language*, 1–52. Cambridge, MA: MIT Press.

Clarke, E., Reichard, U. H., and Zuberbühler, K. (2006). 'The syntax and meaning of wild gibbon songs', *PLoS ONE* 1: e73.

Clay, Z., Smith, C. L., and Blumstein, D. T. (2012). 'Food-associated vocalizations in mammals and birds: What do these calls really mean?', *Animal Behaviour* 83: 323–330.

Coye, C., Ouattara, K., Zuberbühler, K., and Lemasson, A. (2015). 'Suffixation influences receivers' behaviour in non-human primates', *Proceedings of the Royal Society B* 282: 20150265.

Engesser, S. and Townsend, S. W. (2019). 'Combinatoriality in the vocal systems of nonhuman animals', *WIREs Cognitive Science* 10: e1493.

Engesser, S., Ridley, A. R., and Townsend, S. W. (2016). 'Meaningful call combinations and compositional processing in the southern pied babbler', *Proceedings of the National Academy of Sciences of the United States of America* 113: 5976–5981.

Freeberg, T. M. (2008). 'Complexity in the chick-a-dee call of Carolina chickadees (*Poecile carolinensis*): Associations of context and signaler behavior to call structure', *Auk* 125: 896–907.

Freeberg, T. M. and Mahurin, E. J. (2013). 'Variation in note composition of chick-a-dee calls is associated with signaler flight in Carolina chickadees, *Poecile carolinensis*', *Ethology* 119: 1086–1095.

Fujita, K. (2014). 'Recursive Merge and human language evolution', in T. Roeper and M. Speas (eds.), *Recursion: Complexity in Cognition*, 243–264. New York: Springer.

Ghani, B., Klinck, H., and Hallerberg, S. (2023). 'Classification of group-specific variations in songs within House Wren species using machine learning models', *Ecological Information* 74: 101946.

Girard-Buttoz, C., Bortolato, T., Laporte, M., et al. (2022). 'Population-specific call order in chimpanzee greeting vocal sequences', *iScience* 25: 104851.

9780192886491_Book.indb 604 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:07

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Hurford, J. R. (2011). *The Origins of Grammar: Language in the Light of Evolution II*. New York: Oxford University Press.

Janik, V. M. and Knörnschild, M. (2021). 'Vocal production learning in mammals revisited', *Philosophical Transactions of the Royal Society B* 376: 20200244.

Kroodsma, D. E., Byers, B. E., Goodale, E., et al. (2001). 'Pseudoreplication in playback experiments, revisited a decade later', *Animal Behaviour* 61: 1029–1033.

Leroux, M., Schel, A. M., Wilke. C., et al. (2023). 'Call combinations and compositional processing in wild chimpanzees', *Nature Communications* 14: 2225.

Magrath, R. D., Haff, T. M, Fallow, P. M., and Radford, A. N. (2015). 'Eavesdropping on heterospecific alarm calls: From mechanisms to consequences', *Biology Review* 90: 560–586.

Ouattara, K., Lemasson, A., and Zuberbühler, K. (2009). 'Campbell's monkeys concatenate vocalizations into context-specific call sequences', *Proceedings of the National Academy of Sciences of the United States of America* 106: 22026–22031.

Partee, B., ter Meulen, A., and Wall, R. E. (1990). *Mathematical Methods in Linguistics*. Dordrecht: Kluwer Academic.

Pelletier, F. J. (1994). 'The principle of semantic compositionality', *Topoi* 13: 11–24.

Podos, J., Huber, S. K., and Taft, B. (2004). 'Bird song: The interface of evolution and mechanism', *Annual Review of Ecology, Evolution, and Systematics* 35: 55–87.

Rizzi, L. (2016). 'Monkey morpho-syntax and merge-based systems', *Theoretical Linguistics* 42: 139–145.

Schlenker, P., Chemla, E., and Zuberbühler, K. (2016a). 'What do monkey calls mean?', *Trends in Cognitive Science* 20: 894–904.

Schlenker, P., Chemla, E., Schel, A. M., et al. (2016b). 'Formal monkey linguistics: The debate', *Theoretical Linguistics* 42: 173–201.

Schlenker, P., Coye C., Leroux, M., and Chemla, E. (2023). 'The ABC-D of animal linguistics: Are syntax and compositionality for real?', *Biology Review* 98: 1142–1159.

Seyfarth, R. M., Cheney, D. L., Bergman, T., et al. (2010). 'The central importance of information in studies of animal communication', *Animal Behaviour* 80: 3–8.

Suzuki, T. N. (2012). 'Long-distance calling by the willow tit, *Poecile montanus*, facilitates formation of mixed-species foraging flocks', *Ethology* 118: 10–16.

Suzuki, T. N. (2014). 'Communication about predator type by a bird using discrete, graded and combinatorial variation in alarm calls', *Animal Behaviour* 87: 59–65.

Suzuki, T. N. (2016). 'Semantic communication in birds: Evidence from field research over the past two decades', *Ecological Research* 31: 307–319.

Suzuki, T. N. (2021). 'Animal linguistics: Exploring referentiality and compositionality in bird calls', *Ecological Research* 36: 221–231.

Suzuki, T. N. and Kutsukake, N. (2017). 'Foraging intention affects whether willow tits call to attract members of mixed-species flocks', *Royal Society Open Science* 4: 170222.

Suzuki, T. N. and Matsumoto, Y. K. (2022). 'Experimental evidence for core-Merge in the vocal communication system of a wild passerine', *Nature Communications* 13: 5605.

Suzuki, T. N. and Zuberbühler, K. (2019). 'Animal syntax', *Current Biology* 29: R669–R671.

Suzuki, T. N., Wheatcroft, D., and Griesser, M. (2016). 'Experimental evidence for compositional syntax in bird calls', *Nature Communications* 7: 10986.

Suzuki, T. N., Wheatcroft, D., and Griesser, M. (2017). 'Wild birds use an ordering rule to decode novel call sequences', *Current Biology* 27: 1–6.

9780192886491_Book.indb 605 /12_first_proofs/first_proofs_xml_for_typesetting 27-Mar-25 23:05:07

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

Suzuki, T. N., Wheatcroft, D., and Griesser, M. (2018). 'Call combinations in birds and the evolution of compositional syntax', *PLoS Biology* 16: e2006532.

Suzuki, T. N., Wheatcroft, D., and Griesser, M. (2019). 'The syntax–semantics interface in animal vocal communication', *Philosophical Transactions of the Royal Society B* 375: 20180405.

Williams, H. and Lachlan, R. F. (2021). 'Evidence for cumulative cultural evolution in bird song', *Philosophical Transactions of the Royal Society B* 377: 20200322.

Zuberbühler, K. (2019). 'Syntax and compositionality in animal communication', *Philosophical Transactions of the Royal Society B* 375: 20190062.

9780192886491_Book.indb 606 /12_first_proofs/first_proofs/xml_for_typesetting 27-Mar-25 23:05:07

C40P1

OXFORD HANDBOOKS IN LINGUISTICS

THE OXFORD HANDBOOK OF AFRICAN AMERICAN LANGUAGE

Edited by Sonja Lanehart

C40P2 C40P3

THE OXFORD HANDBOOK OF AFRICAN LANGUAGES

Edited by Rainer Vossen and Gerrit J. Dimmendaal

C40P4 C40P5 C40P6

THE OXFORD HANDBOOK OF APPLIED LINGUISTICS

Second edition

Edited by Robert B. Kaplan

C40P7 C40P8 C40P9

THE OXFORD HANDBOOK OF APPROACHES TO LANGUAGE EVOLUTION

Edited by Limor Raviv and Cedric Boeckx

C40P10 C40P11

THE OXFORD HANDBOOK OF ARABIC LINGUISTICS

Edited by Jonathan Owens

C40P12 C40P13 C40P14

THE OXFORD HANDBOOK OF ARCHAEOLOGY AND LANGUAGE

Edited by Martine Robbeets and Mark Hudson

C40P15 C40P16

THE OXFORD HANDBOOK OF CASE

Edited by Andrej Malchukov and Andrew Spencer

C40P17 C40P18

THE OXFORD HANDBOOK OF CHINESE LINGUISTICS

Edited by William S-Y Wang and Chaofen Sun

C40P19 C40P20 C40P21

THE OXFORD HANDBOOK OF COGNITIVE LINGUISTICS

Edited by Dirk Geeraerts and Hubert Cuyckens

C40P22 C40P23

THE OXFORD HANDBOOK OF COMPARATIVE SYNTAX

Edited by Gugliemo Cinque and Richard S. Kayne

C40P24 C40P25

THE OXFORD HANDBOOK OF COMPOSITIONALITY

Edited by Markus Werning, Wolfram Hinzen, and Edouard Machery

C40P26 C40P27

THE OXFORD HANDBOOK OF COMPOUNDING

Edited by Rochelle Lieber and Pavol Štekauer

C40P28 C40P29 C40P30 C40P31

THE OXFORD HANDBOOK OF COMPUTATIONAL LINGUISTICS

Second Edition

Edited by Ruslan Mitkov

C40P32 C40P33

THE OXFORD HANDBOOK OF CONSTRUCTION GRAMMAR

Edited by Thomas Hoffman and Graeme Trousdale

C40P34 C40P35

THE OXFORD HANDBOOK OF CORPUS PHONOLOGY

Edited by Jacques Durand, Ulrike Gut, and Gjert Kristoffersen

C40P36

C40P37 **THE OXFORD HANDBOOK OF DERIVATIONAL MORPHOLOGY**
C40P38 *Edited by Rochelle Lieber and Pavol Štekauer*
C40P39 **THE OXFORD HANDBOOK OF DEVELOPMENTAL LINGUISTICS**
C40P40 *Edited by Jeffrey Lidz, William Snyder, and Joe Pater*
C40P41 **THE OXFORD HANDBOOK OF ELLIPSIS**
C40P42 *Edited by Jeroen van Craenenbroeck and Tanja Temmerman*
C40P43 **THE OXFORD HANDBOOK OF ENDANGERED LANGUAGES**
C40P44 *Edited by Kenneth L. Rehg and Lyle Campbell*
C40P45 **THE OXFORD HANDBOOK OF ENGLISH GRAMMAR**
C40P46 *Edited by Bas Aarts, Jill Bowie, and Gergana Popova*
C40P47 **THE OXFORD HANDBOOK OF ETHIOPIAN LANGUAGES**
C40P48 *Edited by Ronny Meyer, Bedilu Wakjira, and Zelealem Leyew*
C40P49 **THE OXFORD HANDBOOK OF ERGATIVITY**
C40P50 *Edited by Jessica Coon, Diane Massam, and Lisa deMena Travis*
C40P51 **THE OXFORD HANDBOOK OF EVENT STRUCTURE**
C40P52 *Edited by Robert Truswell*
C40P53 **THE OXFORD HANDBOOK OF EVIDENTIALITY**
C40P54 *Edited by Alexandra Y. Aikhenvald*
C40P55 **THE OXFORD HANDBOOK OF EXPERIMENTAL SEMANTICS AND PRAGMATICS**
C40P56 *Edited by Chris Cummins and Napoleon Katsos*
C40P57 **THE OXFORD HANDBOOK OF EXPERIMENTAL SYNTAX**
C40P58 *Edited by Jon Sprouse*
C40P59 **THE OXFORD HANDBOOK OF THE FRENCH LANGUAGE**
C40P60 *Edited by Wendy Ayres-Bennett and Mairi McLaughlin*
C40P61 **THE OXFORD HANDBOOK OF GRAMMATICAL NUMBER**
C40P62 *Edited by Patricia Cabredo Hofherr and Jenny Doetjes*
C40P63 **THE OXFORD HANDBOOK OF GRAMMATICALIZATION**
C40P64 *Edited by Heiko Narrog and Bernd Heine*
C40P65 **THE OXFORD HANDBOOK OF HISTORICAL PHONOLOGY**
C40P66 *Edited by Patrick Honeybone and Joseph Salmons*
C40P67 **THE OXFORD HANDBOOK OF THE HISTORY OF ENGLISH**
C40P68 *Edited by Terttu Nevalainen and Elizabeth Closs Traugott*
C40P69 **THE OXFORD HANDBOOK OF THE HISTORY OF LINGUISTICS**
C40P70 *Edited by Keith Allan*
C40P71 **THE OXFORD HANDBOOK OF INFLECTION**
C40P72 *Edited by Matthew Baerman*
C40P73 **THE OXFORD HANDBOOK OF INFORMATION STRUCTURE**
C40P74 *Edited by Caroline Féry and Shinichiro Ishihara*
C40P75
C40P76
C40P77
C40P78
C40P79

C40P80 **THE OXFORD HANDBOOK OF IRISH ENGLISH**
C40P81 *Edited by Raymond Hickey*
C40P82 **THE OXFORD HANDBOOK OF JAPANESE LINGUISTICS**
C40P83 *Edited by Shigeru Miyagawa and Mamoru Saito*
C40P84 **THE OXFORD HANDBOOK OF LABORATORY PHONOLOGY**
C40P85 *Edited by Abigail C. Cohn, Cécile Fougeron, and Marie Hoffman*
C40P86 **THE OXFORD HANDBOOK OF LANGUAGE AND LAW**
C40P87 *Edited by Peter Tiersma and Lawrence M. Solan*
C40P88 **THE OXFORD HANDBOOK OF LANGUAGE AND PREJUDICE**
C40P89 *Edited by Jane Setter, Sender Dovchin, and Vijay A. Ramjattan*
C40P90 **THE OXFORD HANDBOOK OF LANGUAGE AND RACE**
C40P91 *Edited by H. Samy Alim, Angela Reyes, and Paul V. Kroskrity*
C40P92 **THE OXFORD HANDBOOK OF LANGUAGE AND SOCIETY**
C40P93 *Edited by Ofelia García, Nelson Flores, and Massimiliano Spotti*
C40P94 **THE OXFORD HANDBOOK OF LANGUAGE ATTRITION**
C40P95 *Edited by Monika S. Schmid and Barbara Köpke*
C40P96 **THE OXFORD HANDBOOK OF LANGUAGE CONTACT**
C40P97 *Edited by Anthony P. Grant*
C40P98 **THE OXFORD HANDBOOK OF LANGUAGE EVOLUTION**
C40P99 *Edited by Maggie Tallerman and Kathleen Gibson*
C40P100 **THE OXFORD HANDBOOK OF LANGUAGE POLICY AND PLANNING**
C40P101 *Edited by James W. Tollefson and Miguel Pérez-Milans*
C40P102 **THE OXFORD HANDBOOK OF LANGUAGE PROSODY**
C40P103 *Edited by Carlos Gussenhoven and Aoju Chen*
C40P104 **THE OXFORD HANDBOOK OF LANGUAGES OF THE CAUCASUS**
C40P105 *Edited by Maria Polinsky*
C40P106 **THE OXFORD HANDBOOK OF LEXICOGRAPHY**
C40P107 *Edited by Philip Durkin*
C40P108 **THE OXFORD HANDBOOK OF LINGUISTIC ANALYSIS**
C40P109 *Second edition*
C40P110 *Edited by Bernd Heine and Heiko Narrog*
C40P111 **THE OXFORD HANDBOOK OF LINGUISTIC FIELDWORK**
C40P112 *Edited by Nicholas Thieberger*
C40P113 **THE OXFORD HANDBOOK OF LINGUISTIC INTERFACES**
C40P114 *Edited by Gillian Ramchand and Charles Reiss*
C40P115 **THE OXFORD HANDBOOK OF LINGUISTIC MINIMALISM**
C40P116 *Edited by Cedric Boeckx*
C40P117 **THE OXFORD HANDBOOK OF LINGUISTIC TYPOLOGY**
C40P118 *Edited by Jae Jung Song*
C40P119
C40P120
C40P121
C40P122

C40P123 **THE OXFORD HANDBOOK OF LYING**
Edited by Jörg Meibauer

C40P124 **THE OXFORD HANDBOOK OF THE MENTAL LEXICON**
Edited by Anna Papafragou, John C. Trueswell, and Lila R. Gleitman

C40P125 **THE OXFORD HANDBOOK OF MODALITY AND MOOD**
Edited by Jan Nuyts and Johan van der Auwera

C40P126 **THE OXFORD HANDBOOK OF MORPHOLOGICAL THEORY**
Edited by Jenny Audring and Francesca Masini

C40P127 **THE OXFORD HANDBOOK OF NAMES AND NAMING**
Edited by Carole Hough

C40P128 **THE OXFORD HANDBOOK OF NEGATION**
Edited by Viviane Déprez and M.Teresa Espinal

C40P129 **THE OXFORD HANDBOOK OF NEUROLINGUISTICS**
Edited by Greig I. de Zubicaray and Niels O. Schiller

C40P130 **THE OXFORD HANDBOOK OF PERSIAN LINGUISTICS**
Edited by Anousha Sedighi and Pouneh Shabani-Jadidi

C40P131 **THE OXFORD HANDBOOK OF POLYSYNTHESIS**
Edited by Michael Fortescue, Marianne Mithun, and Nicholas Evans

C40P132 **THE OXFORD HANDBOOK OF PRAGMATICS**
Edited by Yan Huang

C40P133 **THE OXFORD HANDBOOK OF REFERENCE**
Edited by Jeanette Gundel and Barbara Abbott

C40P134 **THE OXFORD HANDBOOK OF RITUAL LANGUAGE**
Edited by David Tavárez

C40P135 **THE OXFORD HANDBOOK OF SOCIOLINGUISTICS**
Second Edition
Edited by Robert Bayley, Richard Cameron, and Ceil Lucas

C40P136 **THE OXFORD HANDBOOK OF SOUTHEAST ASIAN ENGLISHES**
Edited by Andrew J. Moody

C40P137 **THE OXFORD HANDBOOK OF TABOO WORDS AND LANGUAGE**
Edited by Keith Allan

C40P138 **THE OXFORD HANDBOOK OF TENSE AND ASPECT**
Edited by Robert I. Binnick

C40P139 **THE OXFORD HANDBOOK OF THE WORD**
Edited by John R. Taylor

C40P140 **THE OXFORD HANDBOOK OF TRANSLATION STUDIES**
Edited by Kirsten Malmkjaer and Kevin Windle

C40P141 **THE OXFORD HANDBOOK OF UNIVERSAL GRAMMAR**
Edited by Ian Roberts

C40P142
C40P143
C40P144
C40P145
C40P146
C40P147
C40P148
C40P149
C40P150
C40P151
C40P152
C40P153
C40P154
C40P155

THE OXFORD HANDBOOK OF VOWEL HARMONY
Edited by Nancy A. Ritter and Harry van der Hulst

THE OXFORD HANDBOOK OF WORD CLASSES
Edited by Eva van Lier

THE OXFORD HANDBOOK OF WORLD ENGLISHES
Edited by Markku Filppula, Juhani Klemola, and Devyani Sharma

OUP UNCORRECTED PROOF – REVISES, Thu Mar 27 2025, NEWGEN

9780192886491_Book.indb 612 /12_first_proofs/file_typesetting/validation 27-Mar-25 23:05:08