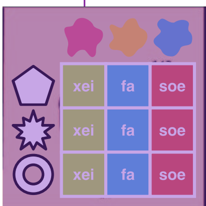
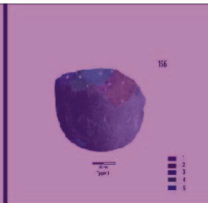
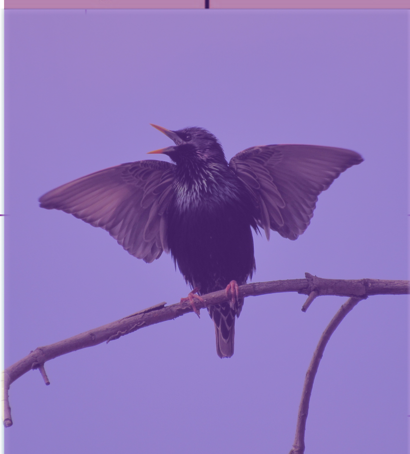
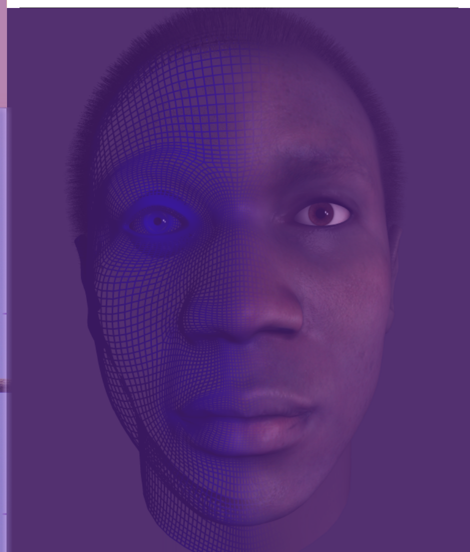
THE EVOLUTION of LANGUAGE

Proceedings of the 15th International Conference









Editors

Jonas Nölle
Limor Raviv
Kirstie E. Graham
Stefan Hartmann
Yannick Jadoul
Mathilde Josserand
Theresa Matzinger
Katie Mudd
Michael Pleyer
Anita Slonimska
Slawomir Wacewicz
Stuart Watson

The Evolution of Language

Proceedings of the 15th International Conference on the Evolution of Language (EVOLANG XV)

Editors:

Jonas Nölle
Limor Raviv
Kirstie Emma Graham
Stefan Hartmann
Yannick Jadoul
Mathilde Josserand
Theresa Matzinger
Katie Mudd
Michael Pleyer
Anita Slonimska
Sławomir Wacewicz
Stuart Watson

EVOLANG XV Scientific Committee
May 18, 2024

Published by:
The Evolution of Language Conferences

Max Planck Institute for Psycholinguistics
Wundtlaan 1
6525 XD Nijmegen
The Netherlands

The Evolution of Language
Proceedings of the 15th International Conference on the Evolution of Language (EVOLANG XV)

ISSN 2666-917X
doi:10.17617/2.3587960

A digital copy of this volume will remain available for download at:
`https://evolang2024.github.io/proceedings/evolang15_proceedings.pdf`

Prefaces

EVOLANG was my introduction to academia. I first attended it in 2002 and the people I met there—and in all the subsequent EVOLANGs—have continued to guide and inspire me. Twenty-two years later, EVOLANG remains my favorite conference.

What makes EVOLANG so good? One answer is that the topic of language origins and evolution is inherently interesting to people with diverse backgrounds and skill sets, making the conference interdisciplinary in the best sense. It bridges disciplines in the pursuit of answers to common questions. But it is more than that. The research community that has formed around EVOLANG encourages both open-mindness and rigor. It takes a broad umbrella approach to incorporating insights from new methods while striving to make continued progress through a collective memory of where the study of language evolution has been and where it is going.

Every EVOLANG I have attended has made me feel the spirit of discovery that I felt that March of 2002. It has been my honor and privilege to share it with all of you as the host of EVOLANG XV in Madison, the city I've called home for the last 14 years.

It is a cliche to say that hosting and running a conference is a team effort. But it's true! I would like to thank Zach Studdiford for technical help and for providing music at the opening reception as the Pathfinder Quartet (a useful reminder of what a difference good music can make!). Maggie Stone provided indispensable help with purchasing, logistics, and excursion-planning, aided by Matt Borman. Eleanor Flannigan and Kate Paape helped solve the problem of inert knowledge. Knowing what needs to be done is only useful if it results in getting things done! Lilia Rissman turned her EVOLANG karaoke dream into reality and hopefully gave many of you a night to remember! I am grateful to the University of Wisconsin Language Sciences Program for their support.

However difficult planning a conference may be, it would have been much more so without the excellent team at Monona Terrace and Destination Madison for their generous grants to promote Madison events.

And, of course, a thank you to my mom, a longstanding honorary EVOLANger, for adding color and wisdom.

See you all in Plovdiv!

May 2024
EVOLANG XV

Gary Lupyan
Professor of Psychology, University of Wisconsin-Madison
Principal local organizer of EVOLANG XV

The 15th International Conference on the Evolution of Language (EVOLANG XV) was held in Madison, Wisconsin on May 18–21, 2024.

As always, the success of the event was facilitated by the team effort and synergy of four bodies: the permanent committee, the local organisers, the scientific committee, and the panel of reviewers. The permanent committee (p.v), headed by Erica Cartmill and Simon Kirby, has always been keen to provide advice and support when needed. Many members of the permanent committee are our mentors or peers, and they keep making the world of language sciences a better place. The local organizers, Gary Lupyan and Robert Hawkins worked hard to make the conference happen. A big thank you to this cohesive team for bringing the 'EVOLANG in Madison' to fruition!

The scientific committee was in charge of editing and reviewing all contributed abstracts and papers, as well as putting the final program together. Once again, EVOLANG XV's scientific committee included a strong involvement of early career researchers in the language evolution community, featuring members at various career stages including PhDs, postdocs and junior PIs from a host of countries and institutions. By building on the expertise of existing members and recruiting new ones, we tried to achieve diversity of scientific backgrounds, covering areas such as developmental psychology, communication, classical linguistics, field research (both in humans and other species), gesture, computational modelling, anthropology, and bioacoustics.

This year, we introduced a new submission and review platform to manage the conference – OpenReview (`https://openreview.net/`). This is a free, open-source and open-access platform that streamlines all the submission and peer-review stages, including editorial assignments, reviewer selection, and announcements of rejection/acceptance. It was definitely a learning process for us (and for all submitters!) seeing as this is quite a novel platform, but overall we feel it was a positive experience and we hope to continue using OR in the next iterations of EVOLANG. We also want to give a special thanks to Yannick Jadoul, who was the technical mastermind behind integrating and managing the OR platform, working night and day to make users' experiences as smooth as possible and solving problems as they arose (and they definitely did!).

We received many high quality submissions, and our reviewers (p.vi) provided important feedback which allowed us to make selections and put together an outstanding program. This volume contains 129 contributions from various disciplines: artificial intelligence, social interaction, syntax, semantics, speech sciences, language acquisition, genetics, bioacoustics, anthropology, animal behaviour, and historical linguistics.

We deeply appreciate the dedication of the local organizing committee, the reviewers, the scientific committee, and the permanent committee, all of whom voluntarily contributed their time. A heartfelt thank you to all fellow editors and committee members for their invaluable support in making this conference a success.

The programme committee members

Jonas Nölle, *Limor Raviv*, *Kirstie Emma Graham*, *Stefan Hartmann*, *Yannick Jadoul*, *Mathilde Josserand*, *Theresa Matzinger*, *Katie Mudd*, *Michael Pleyer*, *Anita Slonimska*, *Sławomir Wacewicz*, *Stuart Watson*

Central Organizing Committee

- Andreas Baumann
- Rudolf Botha
- Christine Cuskley
- Erica Cartmill
- Jean-Louis Dessalles
- Ramon Ferrer i Cancho
- Tecumseh Fitch
- Jim Hurford
- Simon Kirby
- Chris Knight
- Heidi Lyn
- Luke McCrohon
- Kazuo Okanoya
- Thom Scott-Phillips
- Andrea Ravignani
- Nikolaus Ritt
- Kenny Smith
- Maggie Tallerman
- Natalie Uomini
- Sławomir Wacewicz
- Przemysław Żywiczyński

Panel of Reviewers

Ines Adornetti
Marc Allassonnière-Tang
Sarah Babinski
Lluís Barceló-Coblijn
Andreas Baumann
Barend Beekhuizen
Christian Bentz
Antonio Benítez-Burraco
Aleksandrs Berdicevskis
Mélissa Berthet
Arianna Bisazza
Felicia Bisnath
Richard A Blythe
Cedric Boeckx
Irene Boehm
Monika Boruta-Zywiczynska
Alexandra Bosshard
Hannah Brown
Claudio Brozzoli
Olga Capirci
Fausto Carcassi
Jon Carr
Chundra Cathcart
Jules Cauzinille
Chaona Chen
Ching-Yat Cheung
Alexandru Craevschi
Jennifer Culbertson
Christine Cuskley
Aleksandra Cwiek
Emily Davis
Bart de Boer
Marianne De Heer Kloots
Koen de Reus
Valentina Decembrini
Dan Dediu
Peter Dekker
Angelo Damiano Delliponti
Lois Dona
Axel G. Ekström
Caleb Everett
Andrew Feeney
Francesco Ferretti
Vojtěch Fiala
Molly Flaherty
Stefon Flego
Lauren Fletcher
Stella Frank
Michael Franke
Marlen Fröhlich
Susanne Fuchs
Lukas Galke
Rosa Signy Gisladottir
Tao Gong
Kirstie Emma Graham
Kirsty Emma Graham
Oxana Grosseck
Juan Antonio Guerrero Montero
Ankit Gupta
Shreejata Gupta
Felix Haiduk
Harald Hammarström
Stefan Hartmann
Takashi Hashimoto
Robert D. Hawkins
John Hawks
Judith Holler
Annie Holtz
Nathaniel Imel
Rachael E. Jack
Yannick Jadoul
Sverker Johansson
Tamar Johnson
Mathilde Josserand
Gerhard Jäger
Šárka Kadavá
Dariusz Kalociński
Andres Karjus
Vera Kempe
Aislinn Keogh
Say Young Kim
Simon Kirby
Alexander Koplenig
Tom Kouwenhoven
Svetlana Kuleshova
Adriano R Lameira
Imme Lammertink
Dan Lassiter
Maël Leroux
Hannah Little
Gary Lupyan
Hannah Lutzenberger
Heidi Lyn
Calen MacDonald
Vinicius Macuch Silva
Francisco Roman Magdaleno Flores
Alexander Martin
Hiromi Matsumae
Theresa Matzinger
Adrien Meguerditchian
Yasmina Mekki
Jérôme Michaud
Gregory Mills
Steven Moran
Hope E. Morgan
Olivier Morin
Katie Mudd
Elliot Murphy
Thomas Franz Müller
Bill Noble
Jonas Nölle
Asli Ozyurek
Giulia Palazzolo
Elizabeth Pankratz
Joel Parthemore
Louise Peckre
François Pellegrino
Marcus Perlman
Simone Pika
Marek Placiński
Michael Pleyer
Piotr Podlipniak
Wim Pouw
Ljiljana Progovac
Marlou Rasenberg
Heikki Rasilo
Limor Raviv
Casey M Riedmann
Nikolaus Ritt
Gareth Roberts
Sean G Roberts
Carmen Saldana
Sebastian Sauppe
Isaac Schamberg
Adam Schembri
Dominik Schlechtweg
Marieke Schouwstra
Thom Scott-Phillips
Helen Sims-Williams
Anita Slonimska
Kenny Smith
Nicole Tamer
Peeter Tinits
Oksana Tkachman
Rory Turnbull
Kristian Tylén
Slawomir Wacewicz
Stuart Watson
Stuart Kyle Watson
Andrew B Wedel
Franziska Wegdell
Raphael Werner
Lucie Wolters
Noga Zaslavsky
Elizabeth Qing Zhang

Contents

Plenary Talks

The context of transmission matters in the creation and evolution of language

Marie Coppola

marie.coppola@uconn.edu
Department of Psychological Sciences, University of Connecticut, Storrs, Connecticut, USA

In recent decades the study of new deaf communities and their languages has offered us an opportunity to observe language evolution in action. Consistent with laboratory-based studies of artificial language development and change, this work makes clear that transmission is a key part of this process. However, what those changes are, the specific nature of the adaptation and change, as well as the degree to which conventionalization takes place, are influenced by a variety of factors. Major dimensions of transmission that have been examined closely, especially in the context of deaf and hearing signers in Nicaragua, are vertical and horizontal interaction, as well as the quantity and density of connections among signers. Transmission offers opportunities to observe how learners might change the input they receive in the language they produce; our findings thus far suggest that that adaptation is asymmetric (younger learners adapt more, but often in a direction away from their models) and it is also not random. However, the direction it is nudged in is shaped by a variety of factors, including the number of users, their age, the proportion who are primary users of the language, the rate at which new signers are added, and the frequency of interactions (e.g., Senghas, 2005, LeGuen et al., 2020).

I will discuss examples of specific developments in the domains of lexical conventionalization, grammatical uses of space, and pragmatic understanding in Nicaraguan signing. In each case, the nature of the development was partially determined by the context of transmission of the language system. For example, first-cohort signers' use of spatial grammar was transformed differently through vertical transmission to second-cohort signers (who are in horizontal contact) than vertical transmission to their own CODA children (who lack horizontal contact).

Homesigners, who do not enter a linguistic community, consequently lack both vertical and horizontal contact, which reduces opportunities for adaptation and change, potentially resulting in lower levels of convergence on common structures. Such differences reveal that transmission is crucial, not as a source of linguistic content or structure, but as a mechanism that enables language systems to adapt and change. The changes themselves are dynamic responses to the context in which the transmission takes place.

Global and local studies of genetic and linguistic evolution

Nicole Creanza

nicole.creanza@vanderbilt.edu
Department of Biological Sciences, Vanderbilt University, Nashville, Tennessee, USA

Worldwide patterns of genetic variation are driven by human demographic history. In several contexts, we have tested whether this demographic history has left similar signatures on languages to those it has left on genes. Globally, we found a geographic pattern in which populations that were closer to one another tended to be more similar genetically and linguistically. Our analyses suggested that two processes influence this pattern: vertical transmission of both genes and languages during the peopling of the world, and linguistic borrowing (often coupled with genetic admixture) when neighboring populations come into contact, even when their languages are very different. We then build on these findings in multiple contexts. We explore whether sex-biased patterns in human history affect genetic and cultural evolution, by merging genetic, linguistic, and ethnographic data to study the potentially differing signatures of maternal and paternal transmission. We conduct in-depth genetic and linguistic analyses of dialect-level variation within England, and we use similar data to better understand the formation of Sranan, a Creole language in Suriname. Finally, we examine features of Creole languages more broadly to understand how they form and evolve, as well as whether these patterns are reflected in signatures of genetic admixture. Jointly studying linguistic and genetic variation at multiple levels can give us a more nuanced understanding of human evolution and diversity.

The enchronic envelope: A privileged locus in the life cycles of language

Nick Enfield

Nick.Enfield@sydney.edu.au
University of Sydney, Sydney, Australia

A language is a complex adaptive system, with diverse causal processes interacting at different timescales. The causal/temporal frames range from phylogenetic to ontogenetic, microgenetic, enchronic, and diachronic. I argue that these processes converge and interface at a single privileged locus, a 2½-s opportunity for action, called the enchronic envelope. I build on two key claims: (1) language is a form of action and will therefore be structured similarly to physical actions; (2) linguistic actions (like any communicative actions) are subject to a legibility criterion, which strongly constrains the design of linguistic structures in social interaction. I argue that this envelope is where processes at diverse time scales must be realised, including individual-level language learning and population-level conventionalization and change. I seek to focus the sometimes-diffuse idea of language as a complex-system by focusing on a central causal interface for processing, learning, transmitting, and conventionalizing linguistic systems.

Language vitality: Understudied in evolutionary linguistics

Salikoko S. Mufwene

s-mufwene@uchicago.edu
Department of Linguistics, University of Chicago, Chicago, USA

Evolutionary linguistics is not just about the evolution of structures, it is also about *language vitality*, an umbrella term about language birth, vibrance, resilience, endangerment, and death. It conjures up population movements and language contact within specific population structures, with the latter rolling the dice not only on how forms and structures are selected from a joint feature pool into and out of new language varieties but also on which languages prevail, remain vibrant or resilient, or vanish in particular social ecologies.

In the face of linguistic diversity, we probably will never know how many languages were spoken among humans by the time of the Exodus out of Africa. On the other hand, the known history of human migrations, including imperial/colonial expansions, has made obvious that language birth and death have occurred repeatedly. This history makes it imperative for us to investigate the ecological conditions under which these processes have occurred and to assess the current claim that the number of languages has been dramatically decreasing compared to earlier stages of human history.

Thinking of languages as technologies can we explain why some populations give up their languages and whether users or the situations leading them to such shifts should be blamed or pitied for doing so? Is language shift maladaptive? Is it different from other kinds of technology shifts, including religion, health practices, diets, and clothing, among a host of other folk technologies?

Silent gesture: Uncovering biases that shape linguistic conventions

Marieke Schouwstra

M.Schouwstra@uva.nl
Institute for Logic, Language and Computation, University of Amsterdam, Amsterdam, the Netherlands

Everyone likes to be understood. When two individuals do not share a language but still need to communicate, they readily do what they can to bridge the language gap. This adaptability of human communication has made it possible for evolutionary linguists to conduct silent gesture experiments (in which lab participants use their hands and bodies - but no speech - to convey information). Silent gesture experiments have generated valuable insights about the biases and preferences that shape language structure in situations devoid of linguistic conventions. Combining silent gesture with repeated interaction and learning has enabled us to paint an increasingly detailed picture of the forces at play in language emergence, in the lexical as well as the syntactic domain. I will discuss key experiments and the gestural languages they brought forth, and compare them to structures observed in natural languages (new and old, spoken and signed). From this, a picture emerges of language as being shaped by 'lazy' as well as 'zealous' preferences: when creating utterances, language users like to re-use structures they have seen before (in other utterances, or in the world), but while they do so, they actively take into account the shared knowledge with their communication partners – because they like to be understood.

Refereed Contributions

Compression in killer whale pulsed calls

Javier Almunia[1], Jonas Philipp Luke[2], Fernando Luis Rosa González[2], and Ramon Ferrer-i-Cancho[*3]

[*]Corresponding author: rferrericancho@cs.upc.edu
[1]Loro Parque Foundation, Puerto de la Cruz, Spain
[2]Department of Industrial Engineering, Universidad de La Laguna, La Laguna, Spain
[3]Department of Computer Science, Universitat Politècnica de Catalunya, Barcelona, Catalonia, Spain

Decades of research in quantitative linguistics have unveiled that, in spite of the radical differences between languages spoken on Earth (Blasi, Henrich, Adamou, Kemmerer, & Majid, 2022), languages share general statistical patterns called linguistic laws (Zipf, 1949; Altmann, 1980; Bentz & Ferrer-i-Cancho, 2016). A robust instance is Zipf's law of abbreviation, namely, the tendency of more frequent words to be shorter. This law pervades languages independently of many relevant parameters: linguistic family, writing system or measurement unit (Bentz & Ferrer-i-Cancho, 2016; Petrini et al., 2023a). This and other linguistic laws have been found in a wide range of other species (Semple, Ferrer-i-Cancho, & Gustison, 2022). From a theoretical standpoint, these laws are seen as manifestations of principles of communication since Zipf's pioneering research. The law of abbreviation is a prediction of the principle of compression, namely pressure to reduce the magnitude (length or duration) of types (vocalizations or gestures) (Ferrer-i-Cancho et al., 2013; Ferrer-i-Cancho, Bentz, & Seguin, 2022).

Linguistic laws and their underlying principles remain underexplored in cetaceans, who communicate mainly through clicks, whistles and pulsed calls (Dudzinski & Hill, 2017). Evidence of Zipf's rank-frequency law has been reported for dolphin vocalizations (Markov & Ostrovskaya, 1990) and whistles (McCowan, Hanser, & Doyle, 1999). The law of abbreviation, Menzerath's law (longer linguistic constructs tend to be made of smaller parts) and Zipfian laws of word meaning (more frequent words tend to have more meanings) have been reported for dolphin whistles (Vradi, 2021; Ferrer-i-Cancho & McCowan, 2009). Here we expand this research program by adding killer whales, who regulate group movements and cohesion via acoustic communication and exhibit "dialects" that are culturally transmitted (Filatova et al., 2012).

As the view that languages are shaped by cost-cutting considerations is becoming popular (Gibson et al., 2019), recent research has quantified the actual degree of optimization of languages using two variables: the distance between syntactically related words (Ferrer-i-Cancho, Lusseau, & McCowan, 2022) and word lengths (Petrini et al., 2023b; Pimentel, Nikkarinen, Mahowald, Cotterell, &

Blasi, 2021; Ferrer-i-Cancho & Bentz, 2018).

Here we aim test for the presence of Zipf's law of abbreviation in killer whales and also to evaluate, for the 1st time in a non-human species, the degree of optimality of their vocalizations and its temporal evolution by means of a novel optimality score, Ψ, that measures the percentage of optimization of a system: 0% in case of a system that maps type frequencies into type lengths arbitrarily; 100% in case of an optimal coding system (Petrini et al., 2023b). To that aim, we use a dataset of spontaneous pulsed calls produced between 2007 and 2013 by six captive killer whales living in the Loro Parque facilities (Canary Islands, Spain).

We find a significant negative correlation between the frequency of a call type and its duration, in agreement with Zipf's law of abbreviation. To understand the strength of the finding, we also restrict the analysis to specific years. Then the correlation is only significant in 2013. However, three findings support some effect of compression on individual years: (a) the negative correlation that is predicted by optimal coding (Ferrer-i-Cancho et al., 2022) is found in all years except one (2010), (b) the mean duration of call types is below a novel random baseline (Petrini et al., 2023b) for all years and (c) crucially, the sum of the correlations that are obtained over all years is significantly low.

The Ψ score indicates that pulsed calls are optimized to a 38%. This is a rather low degree of optimization compared to word durations in human languages: only Vietnamese, with $\Psi = 33\%$, exhibits a degree of optimality smaller than that of killer whales according to a recent study covering 46 languages (12 families, two constructed languages and one isolate; see results on Common Voice in Petrini et al. (2023b)). That indicates that the duration of killer whale pulsed calls has a degree of optimisation lower than most human languages.

Now we turn our attention to the evolution of call durations. In human languages, there is evidence that orthographic word lengths have been increasing over time (Chen, Liang, & Liu, 2015; Milička, 2018). In killer whales, we do not find any monotonic temporal trend, neither towards longer calls nor towards shorter calls over successive years or months. No monotonic trend is found for the optimality of their duration either. That suggests that, globally, vocalizers have neither increased nor decreased the duration of calls or its optimality in a way that changes persist over time.

To sum up, we conclude that coding efficiency is a property shared not only by humans and a long list of primates (see Safryghin et al. (2022)) but also cetaceans (dolphins and here killer whales). Our findings support the hypothesis that species with distant common ancestors may have converged to the law of abbreviation through the action of the principle of compression. Concerning compression in killer whale pulsed calls, we conclude that (a) the principle is acting with less intensity than in most human languages and (b) its intensity has neither decayed nor increased within the small group and small evolutionary scale (a period of 7 years) we have examined.

Acknowledgements

RFC is supported by a recognition 2021SGR-Cat (01266 LQMC) from AGAUR (Generalitat de Catalunya) and the grants AGRUPS-2022 and AGRUPS-2023 from Universitat Politècnica de Catalunya. The authors thank Isa Bregante, Jasmin Cirilo, Marco Fratini, Trine Hansen, Chiara Ivaldi, Charlotte Kirschner, William Kreikjes, Dorothee Kremers, Estela Lalueza, Yvonne Montenegro, Hector Morales, Marta Ojeda, Monica Vega and David Verchill for annotating parts of the dataset.

References

Altmann, G. (1980). Prolegomena to Menzerath's law. *Glottometrika*, *2*, 1-10.

Bentz, C., & Ferrer-i-Cancho, R. (2016). Zipf's law of abbreviation as a language universal. In C. Bentz, G. Jäger, & I. Yanovich (Eds.), *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics.* University of Tübingen.

Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D., & Majid, A. (2022). Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences*, *26*(12), 1153-1170.

Chen, H., Liang, J., & Liu, H. (2015). How does word length evolve in written Chinese? *PLOS ONE*, *10*(9), 1-12.

Dudzinski, K. M., & Hill, H. M. (2017). Cetacean communication. In J. Vonk & T. Shackelford (Eds.), *Encyclopedia of animal cognition and behavior* (pp. 1–12). Cham: Springer International Publishing.

Ferrer-i-Cancho, R., & Bentz, C. (2018). The evolution of optimized language in the light of standard information theory. In C. Cuskley, M. Flaherty, H. Little, L. McCrohon, A. Ravignani, & T. Verhoef (Eds.), *The evolution of language: Proceedings of the 12th international conference (evolangxii).* NCU Press.

Ferrer-i-Cancho, R., Bentz, C., & Seguin, C. (2022). Optimal coding and the origins of Zipfian laws. *Journal of Quantitative Linguistics*, *29*(2), 165-194.

Ferrer-i-Cancho, R., Hernández-Fernández, A., Lusseau, D., Agoramoorthy, G., Hsu, M. J., & Semple, S. (2013). Compression as a universal principle of animal behavior. *Cognitive Science*, *37*(8), 1565-1578.

Ferrer-i-Cancho, R., Lusseau, D., & McCowan, B. (2022). Parallels of human language in the behavior of bottlenose dolphins. *Linguistic Frontiers*, *5*(1), 5-11.

Ferrer-i-Cancho, R., & McCowan, B. (2009). A law of word meaning in dolphin whistle types. *Entropy*, *11*(4), 688-701.

Filatova, O. A., Deecke, V. B., Ford, J. K., Matkin, C. O., Barrett-Lennard, L. G., Guzeev, M. A., Burdin, A. M., & Hoyt, E. (2012). Call diversity in the

North Pacific killer whale populations: implications for dialect evolution and population history. *Animal Behaviour*, *83*(3), 595-603.

Gibson, E., Futrell, R., Piantadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*, *23*, 389-407.

Markov, V. I., & Ostrovskaya, V. M. (1990). Organization of communication system in Tursiops truncatus Montagu. In J. A. Thomas & R. A. Kastelein (Eds.), *Sensory abilities of cetaceans - laboratory and field evidence* (Vol. 196, p. 599-602). New York: Plenum Press.

McCowan, B., Hanser, S. F., & Doyle, L. R. (1999). Quantitative tools for comparing animal communication systems: information theory applied to bottlenose dolphin whistle repertoires. *Animal Behaviour*, *57*, 409-419.

Milička, J. (2018). Average word length from the diachronic perspective: The case of Arabic. *Linguistic Frontiers*, *1*(2), 81–89.

Petrini, S., Casas-i-Muñoz, A., Cluet-i-Martinell, J., Wang, M., Bentz, C., & Ferrer-i-Cancho, R. (2023a). Direct and indirect evidence of compression of word lengths. Zip's law of abbreviation revisited. *Glottometrics*, *54*, 58-87.

Petrini, S., Casas-i-Muñoz, A., Cluet-i-Martinell, J., Wang, M., Bentz, C., & Ferrer-i-Cancho, R. (2023b). The optimality of word lengths. Theoretical foundations and an empirical study.

Pimentel, T., Nikkarinen, I., Mahowald, K., Cotterell, R., & Blasi, D. (2021). How (non-)optimal is the lexicon? In *North American Chapter of the Association for Computational Linguistics* (pp. 4426–4438).

Safryghin, A., Cross, C., Fallon, B., Heesen, R., Ferrer-i-Cancho, R., & Hobaiter, C. (2022). Variable expression of linguistic laws in ape gesture: a case study from chimpanzee sexual solicitation. *Royal Society Open Science*, *9*, 9220849.

Semple, S., Ferrer-i-Cancho, R., & Gustison, M. (2022). Linguistic laws in biology. *Trends in Ecology and Evolution*, *37*(1), 53-66.

Vradi, A. (2021). *Dolphin communication: a quantitative linguistics approach.* Unpublished master's thesis, Barcelona School of Informatics, Barcelona.

Zipf, G. K. (1949). *Human behaviour and the principle of least effort.* Cambridge (MA), USA: Addison-Wesley.

The emergence of non-absolute synonymy: An iterated-learning experiment

Aja Altenhof[1] and Gareth Roberts[*2]

[*]Corresponding Author: gareth.roberts@ling.upenn.edu
[1]Department of Linguistics, University of Wisconsin-Madison, Madison, USA
[2]Department of Linguistics, University of Pennsylvania, Philadelphia, USA

Synonymy is common, but absolute synonymy—where synonyms can be substituted for one another in any context with no change to truth value, communicative impact, or connotational meaning—has long been recognized to be extremely rare (Cruse, 1986). Why should this be?

One possibility is that this pattern is driven by a cognitive bias against treating words as perfectly synonymous. For example, following the mutual-exclusivity bias, learners assume a single object has a single label (Markman & Wachtel, 1988; Lewis, Cristiano, Lake, Kwan, & Frank, 2020). Alternatively, a lack of absolute synonymy could be driven by accidental differences in the distribution of competing items. Potential synonyms—particularly those arising through borrowing—are rarely entirely equivalent in their sociocultural distributions (cf. Andersen, Furiassi, Mišić Ilić, et al., 2017), which could to lead to them acquiring connotational differences. Over time, these differences could become amplified and lexicalized during learning, pushing synonyms apart.

Altenhof and Roberts (2023) investigated this by exposing participants to two "new slang" verbs in English—*snater* and *fincur*—informing them that the two words had the same meaning, which participants had to guess. The words were presented embedded in English sentences, whose valence was manipulated to imply a negative, positive, or neutral meaning. Next, participants were asked to insert the words into unseen sentences that also differed in terms of valence. A distractor noun (*murp*) was included in both exposure and generalization to reduce demand characteristics. The distribution of words across the different valenced contexts during exposure was manipulated. Participants in all conditions treated the words as if they differed in meaning, even when the words had been presented in the same distribution of sentences in exposure. Participants also did not seem to track quantitative distributions in exposure, though qualitative differences (where words were presented in very reliably different contexts) influenced differentiation. Altenhof and Roberts (2023) took these results as potential evidence for a cognitive bias but also noted substantial variation in participants' statistical learning patterns, leading to complex output distributions. What would happen as a result of exposure to these output distributions? Would distributions stabilize over generations (Smith & Wonnacott, 2010) and, if so, would that involve a stable pattern of non-absolute synonymy?

We investigated this by performing an iterated-learning study (Kirby, Griffiths, & Smith, 2014) with 75 participants arranged into 15 diffusion chains of five generations. The first generation of each chain received the same input language: 12 sentences for the distractor noun and 12 for each novel verb (half positive and half negative). To measure the differentiation of each verb during generalization, for each participant, we calculated a differentiation score by dividing the frequency of each verb in its dominant context by its frequency in all contexts and taking the product of the resulting scores. In line with Altenhof and Roberts (2023), we found that differentiation scores increased in the first generation, suggesting that participants were not treating the verbs as synonymous (Fig. 1). However, unlike some previous work on iterated learning (Smith & Wonnacott, 2010; Smith et al., 2017), we did not find increasing stability and reduction of unpredictable variation over generations. In fact, a one-way ANOVA revealed no significant effect of generation on differentiation score, $F(4, 70) = 0.977, p = 0.426$. Participants also exhibited interesting and substantial heterogeneity in their statistical learning patterns.

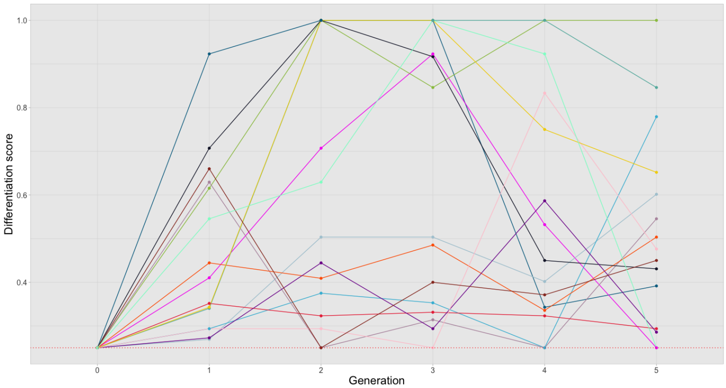


Figure 1. Differentiation scores over all generations.

We discuss these results and their implications alongside ongoing work to replicate the study with modifications designed to control for the role of participant attention and syntactic context. Finally, we discuss the implications of this work for understanding individual differences in statistical learning, an important question for better understanding the cultural evolution of language across generations (Kidd & Arciuli, 2016; Navarro, Perfors, Kary, Brown, & Donkin, 2018).

References

Altenhof, A., & Roberts, G. (2023). Quality, not quantity, impacts the differentiation of near-synonyms. *Language and Cognition*, 1–30.

Andersen, G., Furiassi, C. G., Mišić Ilić, B., et al.. (2017). The pragmatic turn in studies of linguistic borrowing. *Journal of Pragmatics*, *113*, 71–76.

Cruse, D. A. (1986). *Lexical semantics.* Cambridge University Press.

Kidd, E., & Arciuli, J. (2016). Individual differences in statistical learning predict children's comprehension of syntax. *Child development*, *87*(1), 184–193.

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, *28*, 108–114.

Lewis, M., Cristiano, V., Lake, B. M., Kwan, T., & Frank, M. C. (2020). The role of developmental change and linguistic experience in the mutual exclusivity effect. *Cognition*, *198*, 104191.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*(2), 121–157.

Navarro, D. J., Perfors, A., Kary, A., Brown, S. D., & Donkin, C. (2018). When extremists win: Cultural transmission via iterated learning when populations are heterogeneous. *Cognitive Science*, *42*(7), 2108–2149.

Smith, K., Perfors, A., Fehér, O., Samara, A., Swoboda, K., & Wonnacott, E. (2017). Language learning, language use and the evolution of linguistic variation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), 20160051.

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

The relation between European colonialism and linguistic diversity

Andreas Baumann[*1], Bernd Lenzner[2], Hannes Fellner[1], and Franz Essl[2]

[*]Corresponding Author: andreas.baumann@univie.ac.at
[1]Faculty of Philological and Cultural Studies, University of Vienna, Vienna, Austria
[2]Department of Botany and Biodiversity Research, University of Vienna, Vienna, Austria

European colonialism was shown to have entailed a global loss of biodiversity (Crosby, 2004; Yang et al., 2021; Lenzner et al., 2022). Similar effects of colonialism were discussed in the linguistic literature as well (Simons & Lewis, 2013). On the linguistic level, however, the picture seems to be complicated and clearly multicausal (Nettle & Romaine 2000). Most prominently, Mufwene (2002) has suggested colonialism to have differential effects on the linguistic ecosystem, conditioned by the intensity of colonialism in a region. The goal of this paper is to quantitatively examine to what extent the duration under colonial rule in the history of a country is associated with its present status regarding linguistic diversity. In doing so, we examine different operationalizations of linguistic diversity.

Several data resources were combined in this study in order to derive country level measures of colonialism and linguistic diversity. For each country, we used colonization beginning and end dates from COLDAT (Becker, 2019) to estimate 'colonization duration'.[1] The global distribution of colonization time is displayed in Fig. 1. Note that COLDAT is restricted to European empires.

Four ways of measuring linguistic diversity were employed. First, and most straight-forward, we assessed the 'number of languages' currently spoken in each country based on Ethnologue (excluding extinct languages; log-transformed). Second, we computed, for each country, the 'index of linguistic diversity' as introduced by Harmon and Loh (2010). It is computed as one minus the average normalized endangerment level in that country. Endangerment was assessed by means of contemporary EGIDS scores in Ethnologue. Third, we computed 'glottogenetic diversity' as the entropy of

[1] In the case of multiple overlapping colonizers, we only considered the most extreme dates.

language families represented in a country. Glottogenetic information was taken from WALS (Dryer & Haspelmath, 2013). Fourth, we assessed the average 'structural distance' between languages in a country.[2]

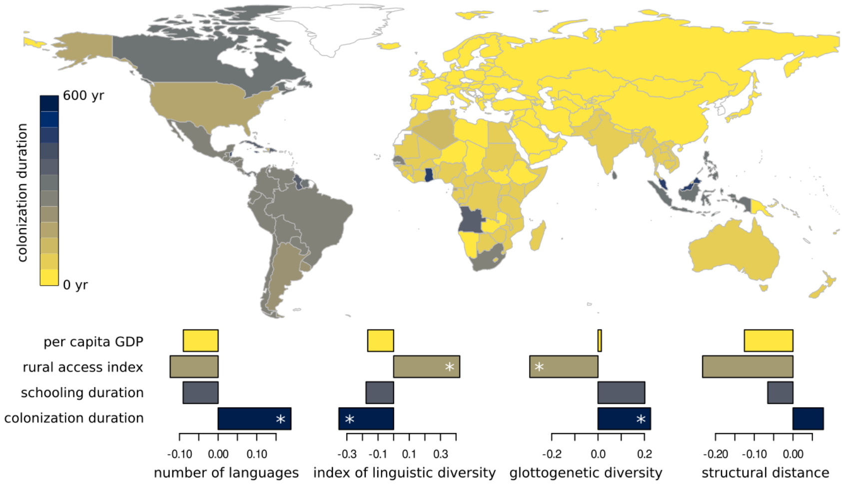


Figure 1. Top: Global distribution of colonization duration. Bottom: standardized coefficients of the linear models ('*' denoting significant effects at a 95% confidence level).

We computed four generalized linear models (GLM, Poisson and quasi-binomial, resp.), one for each measure of linguistic diversity depending on country size (area). In a next step, we computed linear models of the residuals of the four GLMs, featuring colonization duration and a selection of socio-economic covariates (schooling duration, rural access index, per capita GDP; World Bank) that were shown to be relevant to linguistic diversity (Bromham et al., 2022) as predictors (checking for collinearity).

The analysis reveals that colonization duration has differential effects on the four measures of linguistic diversity in the models. More specifically, colonization is negatively related with the linguistic diversity index, but positively with glottogenetic diversity and the number of languages. There is no robust effect on structural distance. One, as we think plausible, interpretation of the results is that while colonization promotes the in-take of genetically distant languages and creolization (Blasi et al. 2017), thereby also increasing glottogenetic diversity, colonization and the implementation of a dominant *lingua franca* has simultaneously lead to an increase in the endangerment of lesser supported ambient languages.

[2] For this, we first computed for each language pair the fraction of non-overlapping linguistic features in Grambank (Skirgård et al. 2023) to obtain pairwise distances. Structural distance was then computed as the mean of all pairwise distances in a country, weighted by the number of features in pairwise comparisons.

References

Becker, B. (2019), Introducing COLDAT: The Colonial Dates Dataset, *SOCIUM/SFB1342 Working Paper Series*, 02/2019.

Blasi, D. E., Michaelis, S. M., & Haspelmath, M. (2017). Grammars are robustly transmitted even during the emergence of creole languages. *Nature Human Behaviour*, 1(10), 723-729.

Bromham, L., Dinnage, R., Skirgård, H., Ritchie, A., Cardillo, M., Meakins, F., Greenhill, S., & Hua, X. (2022). Global predictors of language endangerment and the future of linguistic diversity. *Nature Ecology & Evolution*, 6(2), 163-173.

Dryer, M. S. & Haspelmath, M. (eds.) 2013. *WALS Online (v2020.3)* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7385533

Crosby, A. W. (2004). *Ecological imperialism: the biological expansion of Europe, 900-1900*. Cambridge University Press.

Harmon, D., & Loh, J. (2010). The index of linguistic diversity: A new quantitative measure of trends in the status of the world's languages. *Language Documentation & Conservation*, 4, 97-151.

Lenzner, B., Latombe, G., Schertler, A., Seebens, H., Yang, Q., Winter, M., Weigelt, P., van Kleunen, M., Pyšek, P., Pergl, J., Kreft, H., Dawson, W., Dullinger, S., & Essl, F. (2022). Naturalized alien floras still carry the legacy of European colonialism. *Nature Ecology & Evolution*, 6(11), 1723-1732.

Mufwene, S. (2002). Colonisation, globalisation, and the future of languages in the twenty-first century. *International Journal on Multicultural Societies*, 4(2), 162-193.

Nettle, D., & Romaine, S. (2000). *Vanishing voices: The extinction of the world's languages*. Oxford: Oxford University Press.

Simons, G. F., & Lewis, M. (2013). The world’s languages in crisis. Responses to language endangerment: In honor of Mickey Noonan. *New directions in language documentation and language revitalization*, 3-20.

Skirgård, H., Haynie, H. J., Blasi, D. E., Hammarström, H., Collins, J., Latarche, J. J., Lesage, J., Weber, T., Witzlack-Makarevich, A., Passmore, S., Chira, A., Maurits, L., Dinnage, R., Dunn, M., Reesink, G., Singer, R., Bowern, C., Epps, P., Hill, J., Vesakoski, O., … Gray, R. D. (2023). Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, 9(16), eadg6175.

Yang, Q., Weigelt, P., Fristoe, T. S., Zhang, Z., Kreft, H., Stein, A., ... & van Kleunen, M. (2021). The global loss of floristic uniqueness. *Nature Communications*, 12(1), 7290.

The role of linguistically encoded emotional characteristics for cooperativeness in a one-shot prisoner's dilemma

Andreas Baumann[*1], Theresa Matzinger[2,6], Roland Mühlenbernd[3], Slawomir Wacewicz[5], Michael Pleyer[5], Stefan Hartmann[4], and Marek Placiński[5]

[*]Corresponding Author: andreas.baumann@univie.ac.at
[1]Department of German Studies, University of Vienna, Vienna, Austria
[2]Department of English Studies, University of Vienna, Vienna, Austria
[3]Leibniz-Zentrum Allgemeine Sprachwissenschaft, Berlin, Germany
[4]German Department, Heinrich Heine University Düsseldorf, Düsseldorf, Germany
[5]Center for Language Evolution Studies, Nicolaus Copernicus University in Toruń, Toruń, Poland
[6]Cognitive Science Hub, University of Vienna, Vienna, Austria

We investigate linguistically encoded emotional alignment in pairs of players in a TV game show that is set up as a one-shot prisoner's dilemma. We measure which linguistically encoded emotional characteristics are relevant for choosing between cooperative and defective behavior in that game. We show that cooperativeness depends on interactions between emotional characteristics of both players. In contrast to research on emotional synchrony and cooperation, however, we find that players are more likely to cooperate if their emotions do not align. We interpret this as an instance of deceptive linguistic behavior.

1. Introduction

The cooperative character of language is a key tenet in linguistics, and indeed sharing honest information technically qualifies as cooperation. This presents a well-known evolutionary problem, since, generally, cooperation with biologically unrelated individuals is not evolutionarily stable and can only evolve under very rare circumstances. This is why "the cooperative sharing of information [...] remains a central puzzle in language evolution" (Fitch, 2010: 417). Across the behavioral sciences, the special conditions that enable the emergence and stability of cooperation are typically modeled using the classic game-theoretic tool of the prisoner's dilemma (PD; Nowak & Sigmund, 1993). In this study, we use a PD-structured game show to determine which emotional characteristics may influence the decision to cooperate or to defect.

Previous research indicated that emotions can indeed play a role in maintaining cooperative behavior in PD. Chen et al. (2021) demonstrate that cooperation is promoted in the iterated PD if enough individuals display emotions in a non-competitive way. Similarly, de Melo and Terada (2020) study the effect of non-verbal emotional expression on decision making in the PD.

The alignment of emotions in linguistic interactions was shown to be indicative of cooperation (Arimoto & Okanoya, 2014), and more fundamentally, has been argued to be crucially relevant for the emergence of language in general (Tomasello, 2019). There is robust evidence for emotional alignment and synchrony in parent-child interactions (Lee et al., 2017, Leclère et al., 2014), and among partners (Randall et al., 2013), which are both highly cooperative social relationships. Connected to this, Shilton et al. (2020) argue that emotional synchrony and social bonding are associated and that both have been promoted by coordinated music-making in the social evolution of humans.

Given the close connection of emotion and cooperation, we would expect linguistically encoded emotional alignment to promote cooperativeness, i.e., the tendency to display cooperative behavior in the PD, if players in that game were allowed to communicate before making a decision. This is exactly the hypothesis that we examine in this study. We do so by analyzing linguistically expressed emotional behavior and cooperativeness in a text corpus.

2. Data and preparations

Our study is based on a corpus of 17 transcribed episodes of the TV show 'Golden Balls', a game show that has been the subject of various behavioral studies (e.g., Burton-Chewell & West, 2012). In each episode of this show, four players interact, two of which eventually engage in a final round that effectively represents a one-shot PD, i.e., a variation of the PD in which two individuals play only once. In this game, players can choose to 'split' (cooperate) or 'steal' (defect) the 'jackpot'. The combination of the chosen strategies determines the final reward in line with payoffs in the PD.

In our analysis[1], we only considered utterances from players entering the final round. Since we are interested in emotional characteristics, each utterance was automatically annotated with numeric scores for the following emotional dimensions (Russel & Mehrabian, 1977): valence (V, negative—positive), arousal (A, calm—agitated), and dominance (D, submissive—dominant). We adopted a lexicon-based bag-of-words approach (Taboada et al., 2011) employing VAD norms from Warriner et al. (2013).

[1] Data and code available at https://gitlab.com/andreas.baumann/emo_coop_golden_balls

Next, a smooth time-series model (generalized additive model, Wood, 2017) was fit for each emotional dimension and each player in each episode, thereby describing the trajectory of that emotional property through the episode (Fig. 1, left). Multiple summary measures of the dynamics of VAD of both players were derived from these models: ‘alignedness’ (do the trajectories of both players match?), ‘alignment’ (do the trajectories converge/diverge?), ‘own’ VAD scores, and VAD scores of the ‘other’ player. All measures are listed in Fig. 1 (right).

More specifically, ‘alignedness’ is determined by measuring, for each emotional dimension, dynamic time-warping distance between the trajectories of both players.[2] Low distance, i.e., a high similarity between trajectories, corresponds to high alignedness of both players with respect to that emotional dimension. Measuring emotional ‘alignment’ involves two steps. First, pairwise distances between points on the trajectories for all time-steps (i.e., utterances) in the conversation[3] are computed, i.e., yielding a sequence of distances. Second, a linear regression model is computed in which this distance depends on time. The slope of this model is used for measuring alignment.[4] If the measure is positive, the trajectories of both players start being distant from each other and converge to become more similar in the course of the conversation. If it is negative, the trajectories diverge. In this way, we can differentiate between effects from aligning emotions through the whole conversation and effects of being emotionally synchronized right from the start.

Finally, for each player and each emotional dimension, the ‘own’ value is computed as the average across all scores in the trajectory of that player. The ‘other’ measure is computed, *mutatis mutandis*, by taking the average of all scores of the other player.

3. Importance of emotional features for cooperativeness

To check which measure is most important for predicting player behavior, a linear support vector machine (SVM) with ‘split/steal’ as binary outcome variable was trained and optimized using 5-fold cross-validation. Area under the ROC curve (AUC) was used as a measure of variable importance (Fig 1, right). The model displays an above-chance, albeit not particularly high, accuracy of

---

[2] Dynamic time-warping was chosen to account for potentially shifted emotional reactions in the (pairwise) sequence of utterances.

[3] Note that this is possible since the time-series models interpolate emotion scores so that these models yield predictions for each utterance-step and each player.

[4] Formally, for a linear model of pairwise distance d depending on time t, $d(t) = bt + c + \epsilon$, we define alignment as $-b$. Positive alignment corresponds to convergence, negative alignment to divergence.

0.71 (chance being 0.5).[5] More interestingly, the analysis shows, first, that emotional interactions and the emotions of the other player are considerably more important for behavioral decisions than this is the case for one's own emotions. This is evident since measures of a player's 'own' emotions (valence_own, arousal_own, dominance_own) display low importance. Measures that relate emotions of both players to each other rank higher, on average. Second, we find valence alignedness as well as dominance alignment and alignedness seem to be most important with an AUC score above 0.70, while all other measures are less important (Fig 1, right).

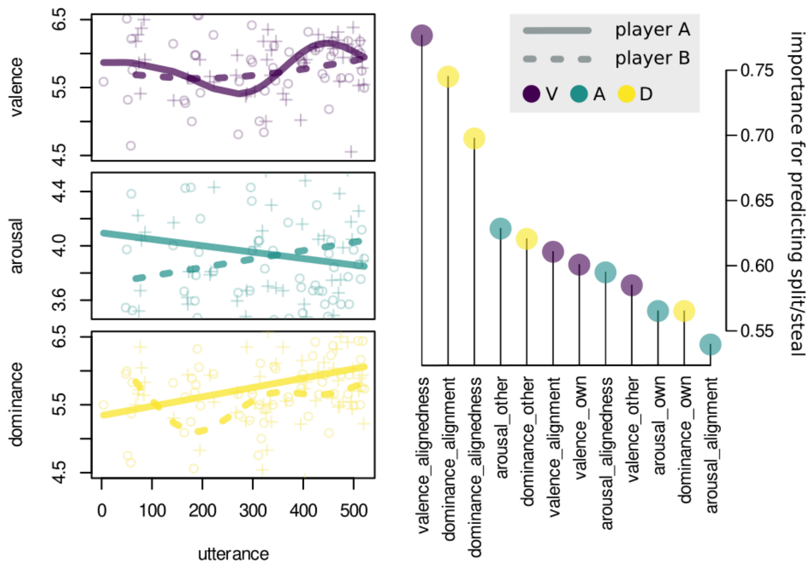


Figure 1. Left: smooth models (GAMs) for emotional developments in one 'Golden Balls' episode. Right: variable importance (ROC AUC) in a SVM based on all episodes.

4. Emotional similarity and cooperativeness

In a second analysis, we tested in more detail how exactly alignment and alignedness influenced cooperation. For each of the three most important predictors of cooperativeness (Fig. 1, right), valence_alignedness, dominance_alignment, dominance_alignedness, we fit a Bayesian Bernoulli model with 'split/steal' as binary outcome variable ('split' being treated as 'success'). We used a logit-link and flat (uninformative) priors for the linear

---

[5] The goal, in the first place, was not to train a model that predicts cooperativeness at a high accuracy, but to gain insights into which (type of) emotional features of a conversation are most relevant for predicting the outcome in an exploratory way. The above chance accuracy at least indicates that the cooperativeness *can* be inferred from emotional characteristics, albeit not reliably.

coefficients. Predictor variables were scaled with respect to their mean and standard deviation before entering the models.

In all models, an effect of emotional alignment/alignedness on the outcome is visible (Fig. 2). However, contrary to our expectations, it is weak rather than strong alignment that promotes one's propensity to cooperate. The respective model coefficients (i.e., effects on the logit) and 95% credible intervals read: -1.27 (-2.36, -0.36) for valence_alignedness, -0.92 (-1.95, -0.07) for dominance_alignedness, and -1.70 (-3.45, -0.34) for dominance_alignment.

What we also see in all models is that low alignment/alignedness yields a chance to split of almost 1.00, while high alignment/alignedness corresponds to a chance to split of around 0.25. Emotional distance seems to be connected to cooperative behavior, while emotional similarity may still entail cooperative behavior at a non-negligible probability.

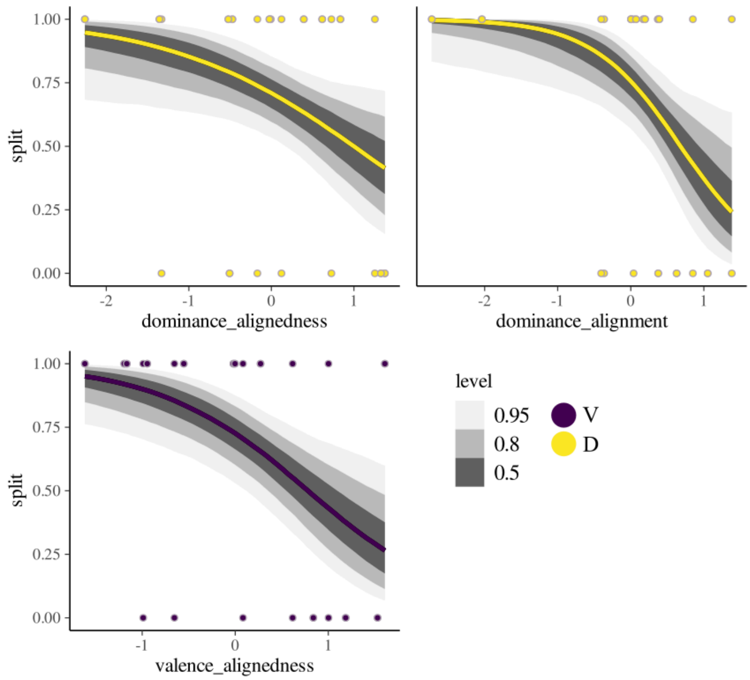


Figure 2. Bernoulli models of cooperativeness (split/steal) depending on the most important variables (cf. Fig. 1, right): dominance_alignedness, dominance_alignment, and valence_alignedness. Bayesian probability bands shown in gray. Emotional closeness generally *decreases* the chance to split.

5. Discussion and conclusion

We have shown that emotional dynamics and interactions among players (rather than just one's own emotions) indeed have an impact on cooperativeness in the

PD, but not in the way that we had expected based on extant research on emotion and cooperation. We found that a player is more likely to cooperate if their counterpart displays *divergent* emotional behavior. Put differently, players are inclined to defect if their emotions are aligned with that of the other player.

This somewhat unexpected outcome could, of course, result from the nature of the data that we inspected. For one, the number of episodes in our sample (17) is relatively small. Although we detect statistically robust effects, it is naturally possible that some of the effects change if more episodes are taken into account. In addition, and more fundamentally, we only assessed emotional expression on the lexical level, thereby ignoring phonetic and prosodic cues, let alone visual information (in particular, gestures or facial expressions; Lei & Gracht, 2019). Finally, the result could be grounded in the artificial setup of the TV show and a potential bias towards competitively minded personalities participating in shows like 'Golden Balls'.

However, leaving the possibility of methodological shortcomings aside, our results could be potentially revealing, as they let us conjecture that linguistically encoded emotion can serve the purpose of deception in competitive situations, thereby also overriding benevolent effects of emotional signaling. That is, emotional alignment could be exploited to mislead a competitor in order to maximize one's own reward. Whether or not this is done consciously cannot be easily assessed based on the examined data.

Interestingly, results from research on emotional mimicry offer an alternative explanation. It was shown that facial mimicry of negative emotions is promoted if one's counterpart has the reputation of behaving in an unfair manner (Hofmann et al., 2012; mimicry of positive emotions was not shown to be modulated by fairness, however). Thus, it could be that players that acquire the reputation of being unfair in the first two rounds of the game and who are expected to defect, elicit (negative) emotional alignment in their counterpart.

In both cases, dishonesty and deception are key aspects. This is in line with the work by Robson (1990) and Santos, Pacheco and Skyrms (2011), who show through evolutionary analyses of the PD with pre-play signaling that signals that are introduced to promote mutual cooperation can easily be exploited towards defection. Moreover, linguistic dishonesty in 'Golden Balls' was already examined in Burton-Chellew and West's (2012) analysis. They found that exaggerating players demoted cooperativeness in their counterpart. Thus, we consider honesty and emotional dynamics in language, and how they impact cooperative behavior to be an interesting interaction worthy of being examined more closely in the light of language evolution research.

Acknowledgements

We would like to thank Jon Carr and an anonymous reviewer for a series of helpful comments on an earlier version of this manuscript. Michael Pleyer was supported by project No. 2021/43/P/HS2/02729 co-funded by the National Science Centre and the European Union Framework Programme for Research and Innovation Horizon 2020 under the Marie Skłodowska-Curie grant agreement No. 945339.

References

Arimoto, Y., & Okanoya, K. (2014). Emotional synchrony and covariation of behavioral/physiological reactions between interlocutors. *17th Oriental Chapter of the International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques (COCOSDA)* (pp. 1-6). IEEE.

Burton-Chellew, M. N., & West, S. A. (2012). Correlates of cooperation in a one-shot high-stakes televised prisoners' dilemma. *PloS one*, 7(4), e33344.

Chen, W., Wang, J., Yu, F., He, J., Xu, W., & Wang, R. (2021). Effects of emotion on the evolution of cooperation in a spatial prisoner's dilemma game. *Applied Mathematics and Computation*, 411, 126497.

Fitch, W.T., 2010. *The Evolution of Language*. Oxford University Press, Oxford.

Hofman, D., Bos, P. A., Schutter, D. J., & van Honk, J. (2012). Fairness modulates non-conscious facial mimicry in women. *Proceedings of the Royal Society B: Biological Sciences*, 279(1742), 3535-3539.

Leclère, C., Viaux, S., Avril, M., Achard, C., Chetouani, M., Missonnier, S., & Cohen, D. (2014). Why synchrony matters during mother-child interactions: a systematic review. *PloS one*, 9(12), e113571.

Lee, T. H., Miernicki, M. E., & Telzer, E. H. (2017). Families that fire together smile together: Resting state connectome similarity and daily emotional synchrony in parent-child dyads. *NeuroImage*, 152, 31-37.

Lei, S., & Gratch, J. (2019). Smiles signal surprise in a social dilemma. *8th International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 627-633). IEEE.

de Melo, C. M., & Terada, K. (2020). The interplay of emotion expressions and strategy in promoting cooperation in the iterated prisoner's dilemma. *Scientific reports*, 10(1), 14959.

Nowak, M., & Sigmund, K. (1993). A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. *Nature*, 364(6432), 56-58.

Randall, A. K., Post, J. H., Reed, R. G., & Butler, E. A. (2013). Cooperating with your romantic partner: Associations with interpersonal emotion coordination. *Journal of Social and Personal Relationships*, 30(8), 1072-1095.

Robson, A. J. (1990). Efficiency in evolutionary games: Darwin, Nash and the secret handshake. Journal of Theoretical Biology, 144, 379–396.

Russell, J. A., & Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. Journal of research in Personality, 11(3), 273-294.

Santos, F.C., Pacheco, J.M., & Skyrms, B. (2011). Co-evolution of pre-play signaling and cooperation. Journal of Theoretical Biology 274, 30–35.

Shilton, D., Breski, M., Dor, D., & Jablonka, E. (2020). Human social evolution: self-domestication or self-control?. *Frontiers in Psychology*, 134.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2), 267-307.

Tomasello, M. (2019). *Becoming human: A theory of ontogeny*. Harvard University Press.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods*, 45, 1191-1207.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. CRC press.

Discursive distinctiveness explains lexical differences between languages

Barend Beekhuizen[1]

[1]Department of Language Studies, University of Toronto Mississauga, Canada
Department of Linguistics, University of Toronto, Canada
barend.beekhuizen@utoronto.ca

Languages differ in their lexical semantic inventories. Recently, such differences have been explored through the lens of differences in the frequencies with which certain concepts are employed in discourse. In this paper, I depart from such work by suggesting that it is not merely the frequency of usage, but also the diversity of ways in which concepts are used that explains whether languages group together two concepts with a single lexical item. I provide a theoretically grounded account of why we should expect this to be the case, and develop a methodology for operationalizing such ideas in a multilingual corpus, finding that variation in the discursive practices of using words indeed predicts whether languages co-express concepts or split them.

1. Introduction

Languages differ in their inventories of lexically encoded meanings. While English co-expresses brothers of both parents as *uncle*, Croatian distinguishes *stric* 'father's brother' from *ujak* 'mother's brother'. Such crosslinguistic variation is the outcome of the cultural evolutionary processes through which only some word meanings are replicated in a community of users. Kemp, Xu, and Regier (2018) explore the usage frequency ('need probability') of concepts as a communicative pressures on the processes of replication: the more often a concept is brought up in discourse, the less likely it is to be co-expressed ('colexified', cf. François, 2008) with similar concepts. This insight has been fruitfully applied to various domains: colour (Twomey, Roberts, Brainard, & Plotkin, 2021), precipitation (Regier, Carstensen, & Kemp, 2016), and kinship (Anand & Regier, 2023).

Here, I propose that rather than the 'need' to express a concept, it is itextbfusage diversity, the diversity of *the ways in which* the concept is employed in discourse that forms a source of selective pressure on word meanings. After presenting the theoretical motivation, I provide support for this position using computational methods, cross-linguistic corpus data, and a lexicon-wide sample of concepts. This result contributes to a more complete account of the pressures shaping the lexicon.

2. Background

The proposed connection between the discursive use of lexical semantic concepts and the inventories of word meanings is motivated by various starting points.

First, lexical selection in usage events is influenced by the inferences afforded by the expressed concepts (Anscombre & Ducrot, 1983; Rommetveit, 1974): we pick lexical items because they steer towards certain conclusions. These inferences do not universally derive from the concept itself (knowing a concept does not entail knowing how it should be used; Goodwin, 1994), but instead depends on semi-conventional 'practices' of speaking (Hanks, 2018) – sets of behavioural patterns governing how a word ought to be used. Such practices of speaking, then, being cultural phenomena, are tethered to a language community, and as such may differ between language communities, as noted by (Hymes, 1961). This motivates the assumption of this paper that the 'rules of use' of lexical items expressing the same or very similar concepts may differ across languages.

Acknowledging a role for language-specific practices of lexical selection is only half of the story. The second half consists of linking those in-the-moment lexical choices to population-level conventions. One proposal to do so comes from Enfield (2014), who takes the in-the-moment decisions, dubbed the 'enchronic' dimension of language, to be one of the 'natural causes' of why language structures are the way they are. Enfield develops a useful conception of how such in-the-moment decisions 'percolate up' to population level conventions in a later paper (Enfield, 2023) in which he argues that part of understanding how concepts are used in discourse is understanding what interpretive effects they have in the past given rise to. Croft (2000) similary takes this 'pool' of experienced usage events to be the source of the selective replication of certain variants over others.

The assumed cultural-evolutionary process for my case is similar. When, in a community, the conventional ways of using two similar concepts are also similar to each other, there is little need to lexically distinguish them, and so new lexical items expressing only one concept are unlikely to emerge and spread. Conversely, when the conventional ways of using the two concepts are different from each other, the concept-level similarity (which might lead to colexification) competes with the dissimilarity on the level of the practice of usage, and we can expect a greater likelihood for e.g., novel lexical items specializing for the expression of one of the concepts to emerge. This paper aims to demonstrate the consequences of these hypothesized pathways for the crosslinguistic patterning of colexification.

3. Method

Studying variation in the discursive usage of word meanings requires a substantially novel set of corpus methods in order to make the crosslinguistic comparison between usage events possible. My method draws on the translation into a shared language (English) to do so. A succinct description is given here, with more information and code made available as part of a planned journal paper.

Corpus: I use the DoReCo corpus (Seifart, Paschen, & Stave, 2022), a typologically diverse sample of fieldwork-based documentation of 51 spoken languages. For comparability, only narrative data was used, resulting in corpora of 500 to 76,000 word tokens per language, with 4 languages excluded for having no narrative data. Around half of the languages have glosses provided for them (e.g., Ex. (1)-(2)), whereas for the remaining languages only the free translation is available (e.g., Ex. (3-4)).

(1)	nam	na	toku	nom	tea	gono	ta	peha	taba	tahii
	1PL.EX.PRON	TAM2	not.know	IPFV	COMPL1	get	NSPEC2.SG	one2	thing	sea

‘we - we don’t know (how) to get anything from the sea.’

(2)	a	abana	paa	nata	vaevuru	tea	vagana
	ART2.SG	men	TAM3	know	already	COMPL1	go.fishing

the men already knew to fish

Teop; Austronesian, Papunesia; (Mosel, 2022)

(3) tayley katiji kastellano
‘I already knew Spanish.’

(4) nish taylejtij
‘We do not know.’

Yurakaré; Isolate, South-America; (Gipper & Ballivián Torrico, 2022)

Extraction of translation equivalents: I use word tokens in the free translations to compare how ‘the same’ concept is expressed across languages. Using SpaCy spacy2, I selected all free translations tokens with ‘lexical’ parts of speech (nouns, adjectives and verbs) and lemmatized them. Next, the most likely orthographic segments and corresponding tokens for each lemma were extracted from the source set using the best-matching string procedure of (Liu et al., 2023). For instance, in Ex. (1) above, for the three lexical items *know*, *get* and *sea*, the Teop strings *toku*, *gono*, and *tahii* were identified as translation equivalents. For the morphologically more complex language Yurakaré (Exx. (3-4)), the English lexical item *know* was linked to the substring *yle* of *tayley*.[1]

Token-level comparability: Massively parallel corpora (e.g., Bible translations) alluw us to compare patterns of colexification through translations of the same source language utterance into all the target languages, but they don’t let us study how concepts are used differently in discourse across languages, as the translations all draw on the same pattern of verbalization in the source language. This motivated the present use of a non-massively parallel corpus that nonetheless has translations *into* a shared target language. To make tokens comparable across languages, I apply computational linguistics techniques for representing the usage of a word through contextualized distributional semantic representations (CDSRs)

[1]The extraction method was found to be highly reliable: evaluating the procedure by considering, for the languages with glosses available, whether the orthographic segment extracted given a free translation matches a target language token glossed with the free translation, we found that the extraction procedure performed at 89% precision and 88% recall (cf. 19% precision/recall if guessing randomly).

in the form of high-dimensional vectors. When tokens of a word are used in similar contexts, their CDSRs will be more similar to each other than when used in different contexts. We expect the CDSRs for *know* in Ex. (1) and Ex. (4) to be similar, as well as those for *knew* in Ex. (2) and Ex. (3), given that each pair represents a similar context. CDSRs for all tokens were retrieved using BERT (`bert-base-cased`; Devlin, Chang, Lee, & Toutanova, 2018).

We can then use the CDSRs to train a supervised classifer to predict the lexical choice in a particular language. Here, I am using the linear Support Vector Machine classifier of `sklearn` (Pedregosa et al., 2011).[2]. A trained classifier allows us to ask, for a token of an English lemma, how that token would be translated in any other language. In other words, we can determine, given the CDSR for *know* in Ex. (4), that the Teop-trained classifier would pick *toku* (as in Ex. (1)), rather than *nata* (as in Ex. (2)), given that the former's contexts are more similar. Doing so for every token and every language, we arrive at a $146,821$-by-47 token-by-language table, where for every token (row) we have the inferred lexical item for each of the target languages (column) in the cells of the table.

Defining lexical fields: To analyze variation in colexification patterns, we need sufficiently large groups of tokens that display crosslinguistic variation. I use the imputed extension of all $9,534$ extracted terms as the starting point, as they reflect groups of tokens colexified by at least one language. I then pairwise merged (by taking the union) term extensions with a Jaccard similarity of $\geq .90$ in order to avoid redundancy (which would affect the regression analyses in Sec. 5), leading to $8,210$ groups of tokens or 'fields'.

Given that the data is a (dummy coded) binary valued table, I ran logistic PCA (Collins, Dasgupta, & Schapire, 2001) to study the patterning of the variation between languages (using the `logisticPCA` library in `R`). Only the first principal component was used as further components might be redundant with the first component of extensions of other terms.

4. A look at the PCA spaces:

As an exploration, I consider a group of tokens colexified by Asimjeeg Datooga (Nilotic, Africa: Griscom, 2022) *nal*, which nearly all translate to English *know*. To understand what the variation along the first component (PC1) of a logistic PCA means, I considered the free translations for the tokens with the lowest and highest value on PC1. The former are overwhelmingly cases of present-tense negated know (e.g., *We don't know because it's a stranger's plan*), whereas the latter consist mainly of instances of past-tense know (e.g., *It was that (which) they knew*). Teop (Mosel, 2022), in Fig. 1a appears to dislexify these two functions,

[2]Classifiers were evaluated using 100-fold cross-validation. The model obtained 87% accuracy in predicting the target language lexical item – outperforming an informed baseline (guessing the most frequent translation equivalent given the English free translation lemma) obtaining 74% accuracy.

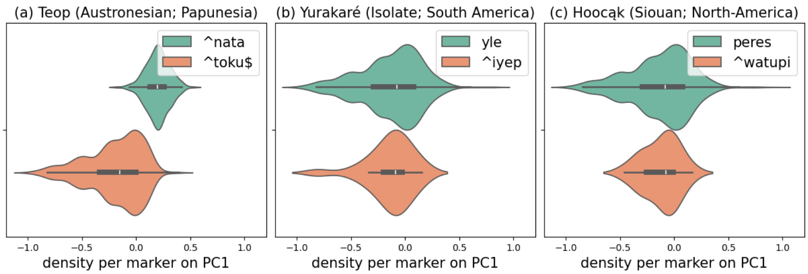


Figure 1. Examples of languages on PC1 of the Asimjeeg Datooga *nal* field.

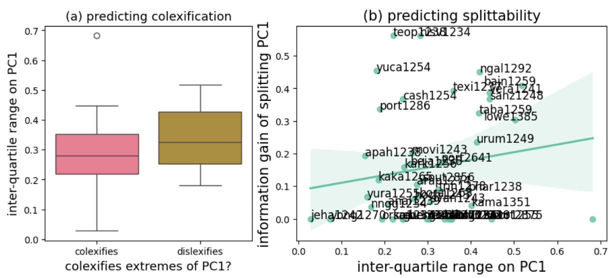


Figure 2. Demonstrating the negative correlation of usage diversity and colexification.

with the two markers seen in the example occupying left and right positions on PC1. Other languages may have multiple terms, but ones that don't line up with the distinction on PC1. Yurakaré (Gipper & Ballivián Torrico, 2022; Fig. 1b), for example, has a second term *iyep* that translates to Spanish *conocer* 'know someone' while *yle* translates to *saber* 'know something', and Hoocąk ((Hartmann, 2022); Fig. 1c), which has, per the provided glosses, a 'know-how' (*watupi*) and a 'know-that' (*peres*) verb. Note that for both languages, the two terms are not linearly separable on PC1.

Notably, the greatest density of Yurakaré and Hoocąk tokens is around the middle of PC1 whereas the tokens of Teop appear to be more spread out. This observation is in line with the central thesis of this paper, that languages that display greater usage diversity colexify less. In particular, I argue that greater variation in the usage tokens on a semantic scale (such as PC1) will go hand in hand with a lower propensity of colexifying the two ends of the scale. Here, I operationalize the usage diversity through the use of the inter-quartile range (IQR) of tokens of a language on the PC1 of a logistic PCA over a group of tokens. As the dependent measure we can consider (1) whether languages would categorize the two extreme points of PC1 with the same term or not (using an SVC trained on the observed markers for each language), and (2) how 'splittable' the language is along the axis. The latter measure makes colexification a continuum by considering the highest information gain (IG) of splitting anywhere along PC1 for a particular language: languages that ‘lump’ will have a zero IG, whereas languages with a 50/50 split between tokens, perfectly splittable along PC1, will have a high IG, and languages with uneven frequency distributions and less-than-perfect splits will fall in between these extremes. Fig. 2a and 2b demonstrate the covariance of two dependent measures covary with usage diversity (IQR) for the *nal* field, showing colexifying languages have a lower IQR than dislexifying ones, and the IG measure correlates positively with the IQR.

5. A lexicon-wide study

Does this correspondence hold in the lexicon at large? I contrast usage diversity (through the IQR) with need probability, defined as the log-transformed word-per-million count of the tokens of a language in the group of tokens considered. Several groups of tokens were omitted for displaying too little variation. This leaves us with 4,679 groups of tokens and 33,843 observations (values for individual languages per field). For our two dependent variables (colexification and splittability) we fit a logistic resp. a linear regression over the two independent variables, z-transforming them for comparability.

For **colexification**, a higher need probability predicts less **colexification** ($\beta = -.08, p < .001$) and a higher usage diversity also predicts less colexification ($\beta = -.72, p < .001$). Both effects are in the expected direction. Moreover, comparing the β values informs us that usage diversity is the more impactful predictor, suggesting that it is not the mere need probability, but the make-up of the discursive need to use a concept that explains differences in colexification. Similarly, we find an effect of usage variation on the **splittability** (information gain) measure in the expected direction ($\beta = .13, p < .001$) but a (smaller) effect for need probability ($\beta = -.02, p < .001$), in the opposite direction, predicting more splittability the less frequent a group of tokens is instantiated for a language.

6. Discussion

This paper studies crosslinguistic variation using naturalistic data for which a substantial methodology had to be developed. The pay-off is that we can study the factors explaining divergence in lexical inventories at scale and using discursive factors that would otherwise not be accessible. The central finding is that crosslinguistic differences in the ways word meanings are used in discourse covary with the types of lexical inventories. This is an initial finding that encourages further consideration of usage events as loci of selectional pressures on the lexicon.

Substantial questions remain, such as the direction of causality between discursive practices and lexical inventories. One could argue that a language having two lexical items nudges their discursive applications to be more distinct. This would not be entirely unexpected, so it is a possibility that we are dealing with a loop, where more discursive distinctiveness drives the need for separate lexicalization, in turn increasing the likelihood of more distinct discourse practices.

Acknowledgements

This research was made possible by an NSERC *Discovery* Grant to the author (RGPIN-2019-06917). I would like to thank audiences at the 2022 International Cognitive Linguistics Conference and the 2022 Societas Linguisticae Europaea for feedback on earlier versions of this work. I am very grateful for the substantial methodological comments presented by the EVOLANG reviewers. I would finally like to express my gratitude to the developers and contributors to the DoReCo corpus, without whom this research would not have been possible.

References

Anand, G., & Regier, T. (2023). Kinship terminologies reflect culture-specific communicative need: Evidence from hindi and english. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 45).

Anscombre, J.-C., & Ducrot, O. (1983). *L'argumentation dans la langue.* Editions Mardaga.

Collins, M., Dasgupta, S., & Schapire, R. E. (2001). A generalization of principal components analysis to the exponential family. *Advances in neural information processing systems*, *14*.

Croft, W. (2000). *Explaining language change: An evolutionary approach.* Pearson Education.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Enfield, N. (2023). Linguistic concepts are self-generating choice architectures. *Philosophical Transactions of the Royal Society B*, *378*(1870), 20210352.

Enfield, N. J. (2014). *Natural causes of language: Frames, biases, and cultural transmission.* Language Science Press.

François, A. (2008). Semantic maps an the typology of colexifications: Intertwining polysemous networks across languages. In M. Vanhove (Ed.), *From polysemy to semantic change: Towards a typology of lexical semantic associations* (pp. 163–216). Amsterdam: John Benjamins.

Gipper, S., & Ballivián Torrico, J. (2022). Yurakaré DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language documentation reference corpus (doreco) 1.2.* Berlin & Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2).

Goodwin, C. (1994). Professional vision. *American anthropologist*, *96*(3), 606–633.

Griscom, R. (2022). Asimjeeg Datooga DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language documentation reference corpus (doreco) 1.2.* Berlin & Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft & labo-

ratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2).
Hanks, W. F. (2018). *Language & communicative practices.* Routledge.
Hartmann, I. (2022). Hooc̨ąk DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language documentation reference corpus (doreco) 1.2.* Berlin & Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2).
Hymes, D. H. (1961). On typology of cognitive styles in language. In *Anthropological linguistics: Uses of typology in language or culture or both: A symposium presented at the 1960 meetings of the american anthropological association* (Vol. 3, pp. 22–54). JSTOR.
Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics*, *4*(1).
Liu, Y., Ye, H., Weissweiler, L., Wicke, P., Pei, R., Zangenfeind, R., & Schütze, H. (2023). A crosslingual investigation of conceptualization in 1335 languages. *arXiv preprint arXiv:2305.08475.*
Mosel, U. (2022). Teop DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language documentation reference corpus (doreco) 1.2.* Berlin & Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2).
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.
Regier, T., Carstensen, A., & Kemp, C. (2016). Languages support efficient communication about the environment: Words for snow revisited. *PloS one*, *11*(4), e0151138.
Rommetveit, R. (1974). *On message structure: A framework for the study of language and communication.* John Wiley & Sons.
Seifart, F., Paschen, L., & Stave, M. (Eds.). (2022). *Language documentation reference corpus (DoReCo) 1.2.* Berlin & Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2).
Twomey, C. R., Roberts, G., Brainard, D. H., & Plotkin, J. B. (2021). What we talk about when we talk about colors. *Proceedings of the National Academy of Sciences*, *118*(39), e2109237118.

Call combinations in bonobos and chimpanzees

Mélissa Berthet[*1,2], Lara Zanutto[1], Morgan Rohée[1], Franziska Wegdell[1,2], Maël Leroux[3], Martin Surbeck[4,5] and Simon W. Townsend[1,2]

[*]Corresponding Author: melissa.berthet.ac@gmail.com

[1]Department of Comparative Language Science, University of Zürich, Switzerland

[2] Department of Evolutionary Anthropology, University of Zürich, Switzerland

[3]EthoS (Éthologie Animale et Humaine) — UMR 6552, University of Rennes I, France

[4]Department of Human Evolutionary Biology, Harvard University, Cambridge, MA, USA

[5]Department of Human Behavior, Ecology and Culture, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

By combining morphemes or words into larger structures, humans can generate an infinite number of meaningful constructions. Despite growing evidence that animals have combinatorial capacities (Berthet et al., 2022), investigation into our closest living relatives, nonhuman apes, remains scarce (Crockford, 2019). Recent observational work (Bortolato et al., 2023; Girard-Buttoz et al., 2022; Leroux et al., 2022) is beginning to address this gap. Specifically, Leroux et al. (2022) studied the chimpanzees *Pan troglodytes* of the Sonso community, at the Budongo Conservation Field Station, Uganda, and identified 15 non-random vocal combinations (Leroux et al., 2022). Here, we followed up on these findings by investigating whether bonobos *Pan paniscus*, the closest living relatives of humans and chimpanzees, also combine calls in systematic ways. We further assessed whether this capacity differs from that of chimpanzees, using data from Leroux et al., 2022.

We conducted 150h of focal recording on 24 adult wild bonobos (14 females and 10 males) from 3 groups at the Kokolopori Bonobo Research Project, Democratic Republic of Congo (Surbeck et al., 2017). During 15-min continuous focal follows, we recorded every vocalization produced by the focal individuals and classified them as one of 10 call types of the bonobo vocal repertoire recently established by Wegdell et al (submitted). An inter-observer reliability test was performed on 10% of the dataset, showing a good agreement between the coder and an external rater (305 calls, K=0.67). Following previous work on great apes (e.g., Leroux et al., 2022), we defined a call combination as two (or more) distinct call types emitted by one individual and separated by less than two seconds of silence.

We collected a total of 1174 utterances comprising 1 to 32 calls (mean=2.64 calls/utterance), and up to 17 call combinations (mean=0.68 combination/utterance). First, we found that the bonobos of Kokolopori vocalize on average 2.5 times more than the Sonso chimpanzees (8.53 vs 3.30 utterances/hour). To specifically investigate their combinatorial capacities, we focused on utterances comprising at least one call combination (N=373 utterances). We used collocation analysis, a method developed in computational linguistics, to detect non-random call combinations, specifically at the bigrammic (i.e., two calls) level (Bosshard et al., 2022). To analyze utterances longer than two calls, we decomposed them into bigrams: for instance, a combination ABC was processed as two separate bigrams AB and BC. A Multiple Distinctive Collocation Analysis (MDCA), showed that, similarly to chimpanzees, bonobos produce several non-random bigrams (N=17). Interestingly, bonobos produce non-random bigrams more frequently than chimpanzees, both in terms of production rate (3.90 vs 0.47 non-random call combinations/hour) but also as a proportion of their total vocal output (non-random bigrams represent 31.9% vs 15.1% of the total vocal production). Additionally, 7 (41%) of the bonobo non-random bigrams are unidirectional (e.g., we observe AB but not BA), suggesting that the order may be important, a finding that closely aligns with that of chimpanzees, where 46% of the non-random bigrams are also unidirectional. Finally, similarly to chimpanzees, male bonobos produce more single utterances than females (16.8 vs 9.3 utterances/hour), but males and females produce combinations at similar rates (4.0 vs 3.8 non-random bigrams/hour respectively). Overall, our results indicate that the vocal communication of bonobos from the Kokolopori population extensively relies on combinations. Moreover, they are more vocal and produce call combinations more frequently than the Sonso community of chimpanzees. We consider a number of social and ecological explanations for these differences. Further investigations should include an evaluation of the meaning of these combinations as well as a replication in other chimpanzee and bonobo communities to assess the more general nature of our findings. Overall, our study provides further tentative support for the hypothesis that the human combinatorial capacity is deeply rooted in the primate lineage.

References

Berthet, M., Coye, C., Dezecache, G., & Kuhn, J. (2022). Animal linguistics: A primer. *Biological Reviews*, brv.12897. https://doi.org/10.1111/brv.12897

Bortolato, T., Friederici, A. D., Girard-Buttoz, C., Wittig, R. M., & Crockford, C. (2023). Chimpanzees show the capacity to communicate about concomitant daily life events. *iScience*, 108090. https://doi.org/10.1016/j.isci.2023.108090

Bosshard, A., Leroux, M., Lester, N. A., Bickel, B., Stoll, S., & Townsend, S. W. (2022). From collocations to call-ocations: Using linguistic methods to quantify animal call combinations. *Behavioral Ecology and Sociobiology Volume*, *76*, 122. https://doi.org/10.1007/s00265-022-03224-3

Crockford, C. (2019). Why does the chimpanzee vocal repertoire remain poorly understood and what can be done about it? In C. Boesch, R. Wittig, C. Crockford, L. Vigilant, T. Deschner, & F. Leendertz (Eds.), *The Chimpanzees of the Taï Forest* (1st ed., pp. 394–409). Cambridge University Press. https://doi.org/10.1017/9781108674218.025

Girard-Buttoz, C., Zaccarella, E., Bortolato, T., Friederici, A. D., Wittig, R. M., & Crockford, C. (2022). Chimpanzees produce diverse vocal sequences with ordered and recombinatorial properties. *Communications Biology*, *5*(1), 410. https://doi.org/10.1038/s42003-022-03350-8

Leroux, M., Chandia, B., Bosshard, A. B., Zuberbühler, K., & Townsend, S. W. (2022). Call combinations in chimpanzees: A social tool? *Behavioral Ecology*, *33*(5), 1036–1043. https://doi.org/10.1093/beheco/arac074

Surbeck, M., Coxe, S. J., & Lokasola, A. L. (2017). Lonoa: The establishment of a permanent field site for behavioural research on bonobos in the Kokolopori Bonobo Reserve. *Pan Africa News*, *24*(2), 13–15.

The study of sign languages and gesture at Evolang conferences

Felicia Bisnath*[1] and Molly Flaherty[2]

*Corresponding Author: fbisnath@umich.edu
[1]Linguistics, University of Michigan, Ann Arbor, MI, USA
[2]Psychology, Davidson College, NC, USA

1. Introduction

What would a naïve attendee at Evolang glean about sign languages and gesture solely based on Evolang presentations? The presence of these topics at Evolang is complex because their clear theoretical and methodological importance (Brentari & Coppola, 2012; Sandler, 2013; Benítez-Burraco, 2015) cannot be divorced from the fact that sign language creators are deaf/hard-of-hearing people, who are dehumanised, deprived of language, and excluded from the academy (Woodcock, Rohan, & Campbell, 2007; Lane, 2017). These facts influence their study, and motivate evaluation of their framing in language evolution. This study explores two aspects of sign and gesture related topics at Evolang: (i) the diversity of sign languages and study methods to identify how language evolution understands manual communication, and (ii) the arguments that sign language and gesture are invoked in, to evaluate the latter's juxtaposition with broader themes in language evolution.

2. Data & Coding

The data set was compiled by filtering abstracts from Evolang 8, 9, 11 and 13, and JCoLE (2022) for the keywords "sign language" and "gesture" occurring in the title. These abstracts were screened manually to find those that substantively dealt with these topics. Our results are based on a subset of the full data set (N=34). The coding categories (Table 1) were developed based on the content of the data set, and implemented by the authors.

3. Findings & Discussion

Items 1-3 below summarise three preliminary results, with codes arranged in descending order of frequency (frequency shown in brackets).

1. Study types: EXPERIMENT[1] (11) > CORPUS WORK (6), NON-HUMAN PRI-

[1]participants are asked to do a task that is not elicitation of a language of which they are a user e.g. silent gesture, artificial language learning, director-matcher, iterated learning.

MATE PARADIGM (6), THEORY (6) > ELICITATION (3) > META-ANALYSIS (2)

2. Sign languages: NICARAGUAN SL (6) > HOMESIGN (4) > KATA KOLOK (1), EMERGING (1), BRITISH SL (1)

3. Themes: CHANGE OVER TIME (11) > CHILD LANGUAGE (7) > GESTURAL ORIGIN (6), LANGUAGE IN CONTEXT (6), NON-HUMAN PRIMATES (6) > EMERGING LANGUAGES (5), ICONICITY (5) > NEURAL UNDERPINNINGS (4) > X CONDITIONING STRUCTURE[2] (3) > BEING HUMAN (2)

The results suggest that understanding of sign languages in language evolution is based on just 3 named sign languages, with studies of Nicaraguan SL prevailing. Data about manual human communication are primarily non-naturalistic apart from corpus work[3]. Thematically, the greatest focus is on change over time[4], primarily examined through age-related comparison, or across experimental transmission chains. Child language as a window onto evolution is a distant second, and tends to be represented by homesign.

Table 1. Coding category definitions

Category	Description
THEME	Frames in language evolution that sign languages and/or gesture are invoked in e.g. gesture is often linked to gestural theories of language evolution
DOMAIN	communicative resources investigated e.g. word order
STUDY TYPE	major method employed e.g. meta-analysis
SIGN LANGUAGE	named sign language or sign language type e.g. Nicaraguan SL, homesign

4. Conclusion

Hammarström (2016) argues for linguistic diversity in language evolution studies. We show that over 12 years, our naïve attendee might have a limited idea of structural and societal diversity in *natural* sign language use, and an idea that the study of manual communication is about grading phenomena as more or less linguistic (Kusters & Sahasrabudhe, 2018; Kusters & Hou, 2020; Kusters, Green, Moriarty, & Snoddon, 2020). This suggests that Evolang should make more active efforts at increasing the diversity of research on manual communication presented at the conference. A step toward doing this that can also identify submission bias is to track the properties of abstracts and assess differences over time.

[2] X = modality, society e.g. community size, semiotic resources.

[3] work from any set of human language data (collected for the study or previously).

[4] We include topics such as transmission, structural reduction, developmental clines, general language change, conventionalisation, emergent systems in the category of change over time.

References

Benítez-Burraco, A. (2015). A biolinguistic approach to sign languages. *The Oxford handbook of deaf studies in language*, 247.

Brentari, D., & Coppola, M. (2012). What sign language creation teaches us about language. *First Language*.

Hammarström, H. (2016). Introduction: Linguistic diversity and language evolution. *Journal of Language Evolution*, 19–29.

Kusters, A., Green, M., Moriarty, E., & Snoddon, K. (2020). Sign language ideologies: Practices and politics. *Sign language ideologies in practice*, 3–22.

Kusters, A., & Hou, L. (2020). Linguistic ethnography and sign language studies. *Sign Language Studies*, *20*(4), 561–571.

Kusters, A., & Sahasrabudhe, S. (2018). Language ideologies on the difference between gesture and sign. *Language & Communication*, *60*, 44–63.

Lane, H. (2017). A chronology of the oppression of sign language in france and the united states. In *Recent perspectives on american sign language* (pp. 119–161). Psychology Press.

Sandler, W. (2013). Vive la différence: Sign language and spoken language in language evolution. *Language and Cognition*.

Woodcock, K., Rohan, M. J., & Campbell, L. (2007). Equitable representation of deaf people in mainstream academia: Why not? *Higher Education*, *53*, 359–379.

Spontaneous emergence of large shared signalling systems with and without referential transmission

Richard A Blythe[*1,2] and Casimir Fisch[1,3]

[*]Corresponding Author: r.a.blythe@ed.ac.uk
[1]School of Physics and Astronomy, University of Edinburgh, Edinburgh, UK
[2]Centre for Language Evolution, University of Edinburgh, Edinburgh, UK
[3]Department of Earth Sciences, ETH Zürich, Zürich, Switzerland

The transition from a state in which signals are used randomly, and are therefore uninformative, to one in which multiple agents use the same signal to convey a specific meaning has long been studied in the field of language evolution (e.g. Hurford, 1989) and related fields (e.g. Lewis, 1969). Although initially most communicative interactions will fail, models (Spike, Stadler, Kirby, & Smith, 2017; Lipowska & Lipowski, 2022; Zubek, Korbak, & Rączaszek-Leonardi, 2022) and experiments (Galantucci & Garrod, 2011) each demonstrate that chance agreement between a pair of agents on a signal's meaning can seed the growth of an *optimal* signalling system in which the probability of a successful communication is maximised. In a recent survey, Spike et al. (2017) propose three basic ingredients that are required to make this work. These are: (i) transmission of referential information; (ii) a bias against ambiguity; and (iii) loss of memory of specific interactions over time. Roughly speaking, these are needed so that (i) the hearer has a chance of guessing the correct meaning; (ii) the system is guided towards optimality; and (iii) a suboptimal state does not become frozen in.

Here, we build on this work by addressing one technical and two conceptual limitations of the many models that Spike et al. (2017) unify into a common framework. The technical limitation is that conclusions and generalisations drawn in the language evolution literature are often based on simulations that are limited in terms of the system size (numbers of agents, meanings and signals) that can be accessed. A natural question is whether arbitrarily large signalling systems (e.g., those with many meanings) can spontaneously arise. At the conceptual level, many studies focus on the ideal case where the number of available signals equals the number of meanings to convey. The question of what happens when there are many more possible meanings than signals available to express them is less well explored. Most fundamentally, it is almost always assumed that agents have some means to communicate whether the intended referent was successfully communicated, whether explicitly (e.g. by pointing, Steels & Belpaeme, 2005) or implicitly through the response to an environmental state that delivers a payoff to signaller and receiver (e.g. Lewis, 1969). One may worry what benefit is conferred by signalling when agents already have available some other reliable means to transmit their referential intention (Oliphant & Batali, 1997).

We address all three limitations by constructing a unified *mathematical* model that includes as special cases many of the different simulation models that have been studied. The general case manifests as a combination of reinforcement learning with memory loss and lateral inhibition (simulated explicitly by Oh & Kim, 2021). It can further be related to the replicator equations of evolutionary game theory (Nowak & Sigmund, 2004), wherein referential transmission enters into the fitness. Particularly, we can apply linear stability analysis (Glendinning, 2012) about an initially uninformative state to identify when multiple agents simultaneously amplify the same signal-meaning associations. This analysis reveals that although increasing the number of agents or number of signals reduces the rate at which associations systematise, this does not pose a barrier to the emergence of an informative signalling system. On the other hand, increasing the size of the meaning space whilst holding the number of signals fixed can render the uninformative initial state inescapable when agents punish a failed interaction by decreasing the relevant meaning-signal association. In short, rewarding success, but ignoring failure, is a robust mechanism for building a shared communication system of arbitrary size.

Most significantly, we find the same outcome is possible even when agents cannot assess (let alone communicate) the success or failure of an interaction, but instead resort to cues (which need not be linguistic) to guess a plausible meaning. This is distinct from the scenario in signalling games (Lewis, 1969; Skyrms, 2010), where the signaller seeks to convey a hidden environmental state to a recipient, whose subsequent behaviour then confirms if their inferrence was correct or not. In our approach, the recipient appeals to *cross-situational learning* (Siskind, 1996), where repeated uses of a word in similar contexts allows a child to reconstruct an adult's pre-existing mapping between words and meanings, even when every instance of use is infinitely ambiguous and there is zero feedback (Blythe, Smith, & Smith, 2016). In the present work, we show that a common lexicon can be built through the same learning mechanism, even when starting from a state in which signals are uninformative. In other words, shared contexts of use are sufficient to provide the referential information required for an optimal signalling system to emerge, despite the presence of ambiguity and no pre-existing means for agents to judge or communicate the success or failure of their interaction.

Taken together, our findings suggest that small-scale computational and laboratory models of the emergence of linguistic systems and structures are representative of what happens in larger and more complex systems. Moreover, they demonstrate the existence of a process by which a species with no pre-existing ability to transmit referential information may acquire the ability to do so.

References

Blythe, R. A., Smith, A. D. M., & Smith, K. (2016). Word learning under infinite uncertainty. *Cognition*, *151*, 18–27.

Galantucci, B., & Garrod, S. (2011). Experimental semiotics: a review. *Frontiers in Human Neuroscience*, *5*, 00011.

Glendinning, P. (2012). *Stability, instability and chaos: An introduction to the theory of nonlinear differential equations.* Cambridge: Cambridge University Press.

Hurford, J. R. (1989). Biological evolution of the saussurean sign as a component of the language acquisition device. *Lingua*, *77*, 187–222.

Lewis, D. (1969). *Convention: A philosophical study.* Oxford: Blackwell.

Lipowska, D., & Lipowski, A. (2022). Emergence and evolution of language in multi-agent systems. *Lingua*, *272*, 103331.

Nowak, M. A., & Sigmund, K. (2004). Evolutionary dynamics of biological games. *Science*, *303*, 793–799.

Oh, P., & Kim, S. (2021). An evolutionary model of the emergence of meanings. *Communication Methods and Measures*, *15*, 255–272.

Oliphant, M., & Batali, J. (1997). Learning and the emergence of coordinated communication. *Center for Research on Language Newsletter*, *11*, 1–46.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*, 39–91.

Skyrms, B. (2010). *Signals: Evolution, learning, and information.* Oxford: Oxford University Press.

Spike, M., Stadler, K., Kirby, S., & Smith, K. (2017). Minimal requirements for the emergence of learned signaling. *Cognitive Science*, *41*, 623–658.

Steels, L., & Belpaeme, T. (2005). Coordinating perceptually grounded categories through language: A case study for colour. *Behavioural and Brain Sciences*, *28*, 469–529.

Zubek, J., Korbak, T., & Rączaszek-Leonardi, J. (2022). *Models of symbol emergence in communication: a conceptual review and a guide for avoiding local minima.* (arxiv:2303.04544)

The Power of Linguistic Similarity for Unlocking Cooperation - Evidence from Syntax and Pitch Experiments

Irene Böhm*[1], Theresa Matzinger[1,2], Marek Placiński[3], Przemysław Żywiczyński[3], and Sławomir Wacewicz[3]

*Corresponding Author: irene.boehm@univie.ac.at
[1]Department of English and American Studies, University of Vienna, Vienna, Austria
[2]Cognitive Science Hub, University of Vienna, Vienna, Austria
[3]Center for Language Evolution Studies, Nicolaus Copernicus University in Toruń, Toruń, Poland

It is important to know who is a good cooperation partner, and current research highlights how language can be a key signal of cooperativeness (Henrich & Henrich, 2007; Matzinger et al., 2023). In particular, low-level linguistic mechanisms such as subconsciously matching others' language have been proposed to be particularly honest signals utilized to assess others' cooperative potential (Wacewicz et al., 2017). One of the big questions at the moment is why these mechanisms are used to select others as cooperation partners. Two possible explanations have been proposed: on the one hand, low-level linguistic similarity (i.e., continuous "*alignedness*" from the start of a conversation) can indicate group members (Dunbar, 1996; Axelrod et al., 2004), and it is known that in-group cooperation is less risky and more successful (e.g. Balliet et al., 2014). On the other hand, adapting to others' linguistic choices (i.e., progressive "*alignment*" throughout a conversation) can indicate others' willingness to cooperate, since it can signal an initial cognitive investment in the cooperation (Kulesza et al., 2014; Chartrand & Bargh, 1999).

To explore how people tend to cooperate with linguistically similar conversation partners, we conducted an experiment on the effect of syntactic similarity on people's choice in cooperation partners (Matzinger et al., 2023). In this picture-description experiment (cf. Bock, 1986), 100 participants communicated with conversation partners, who were in fact bots, that either did or did not match the participants syntactic choices. Based on this language use, the participants then had to decide with whom to cooperate in a subsequent cooperative task. Crucially, half of the participants could freely use their naturally preferred constructions (e.g., “X lends Y to Z”), while the other half were assigned a construction that was not their natural preference (e.g., “X lends Z Y”).

In a logistic regression model, we found that when participants could communicate in their own preferred structures, they predominantly chose linguistically similar conversation partners as cooperation partners (77.0%, 95%-confidence interval [69.0;85.0]). However, when participants were restricted in their language use, they preferred those partners that matched their actual linguistic preference (59.3% 95%-confidence interval [50.2;68.5]), instead of the ones that were similar to their overt linguistic use. We take this to mean that the sheer act of adapting to someone’s linguistic production is not as crucial for choosing cooperation partners, even if it involves an initial investment. Rather, the decisive factor is sharing someone’s linguistic preferences and thereby indicating social group membership. This highlights that the influence of *alignedness* vs. *alignment* needs to be disentangled further in cooperation research.

Therefore, we will expand this research in a follow-up study that hones in on this distinction and tests perceived cooperativeness in a more natural and revised setting. Most importantly, we will focus on pitch instead of syntactic similarity to eliminate the potential confounding factor of priming (Pickering & Garrod, 2004): Alignment does not need to be a conscious investment on the side of the speaker, but can also be a result of purely mechanistic and automatic priming and may, therefore, not be taken as a signal of cooperativeness by the listener. While syntax primarily targets priming, continuous phonological features such as pitch, which are harder to match automatically, have been shown to be less prone to priming and may be a better indicator of active cooperative intentions (Gijssels et al., 2016). Therefore, pitch similarity is particularly well-suited to teasing apart the role of socially-motivated *alignment* vs. *alignedness* in cooperative encounters.

In our talk, we will present theoretical considerations on disentangling the role of *alignedness* and *alignment* for cooperation and set these insights in relation to the results of our study on syntax. We will supplement this with the first findings of our follow-up experiment on the perceived cooperativeness of conversation partners speaking with a pitch that is a) aligned from the start of the conversation, b) aligning throughout the conversation, and c) dissimilar throughout the conversation. In line with the results on syntactic alignment, we predict that interlocutors in group a) will be considered as most cooperative, followed by group b), while group c) will be assessed as least cooperative.

Ultimately, understanding the relationship between language and cooperation in social groups will help us shed light on the evolution and stabilization of both of these traits, which are particularly prominent in humans.

References

Axelrod, R., Hammond, R. A., & Grafen, A. (2004). Altruism via kin-selection strategies that rely on arbitrary tags with which they coevolve. *Evolution, 58*(8), 1833–1838.

Balliet, D., Wu, J., & De Dreu, C. K. (2014). Ingroup favoritism in cooperation: a meta-analysis. *Psychological Bulletin, 140*(6), 1556.

Bock, J. K. (1986). Meaning, sound, and syntax: Lexical priming in sentence production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*(4), 575.

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception–behavior link and social interaction. *Journal of Personality and Social Psychology, 76*(6), 893–910.

Dunbar, R. (1996). *Grooming, Gossip and the Evolution of Language*. Cambridge: Harvard University Press.

Gijssels, T., Casasanto, L. S., Jasmin, K., Hagoort, P., & Casasanto, D. (2016). Speech accommodation without priming: The case of pitch. *Discourse Processes, 53*(4), 233-251.

Henrich, N., & Henrich, J. P. (2007). *Why humans cooperate: A cultural and evolutionary explanation*. Oxford University Press.

Kulesza, W., Dolinski, D., Huisman, A., & Majewski, R. (2014). The echo effect: The power of verbal mimicry to influence prosocial behavior. *Journal of Language and Social Psychology, 33*(2), 183–201.

Matzinger, T., Placiński, M., Lewandowski, M., Gutowski, A., Żywiczyński, P., & Wacewicz, S. (2023). Linguistic preference outcompetes alignment as a predictor for assessing others' cooperativeness. [preprint]

Pickering, M., & Garrod, S. (2004). The interactive-alignment model: developments and refinements. *Behavioral and Brain Sciences, 27*(2), 212-225.

Wacewicz, S., Żywiczyński, P., & Chiera, A. (2017). An evolutionary approach to low-level conversational cooperation. *Language Sciences, 63*, 91–104.

How many words is a picture (or a definition) worth? A distributional perspective on learning new word meanings

Matthew Z. Borman*[1], Gary Lupyan[1]

*Corresponding Author: borman@wisc.edu
[1]Department of Psychology, University of Wisconsin-Madison, Madison, WI, USA

As adults, we continue to learn new word meanings. We can learn new words through ostensive labeling events where a word denotes a clear referent in context, or by having the word explicitly defined for us (Hahn & Gershkoff-Stowe, 2010). However, people also learn word meanings through exposure to how words are used in text (Nagy et al., 1985; Saragi et al., 1978). Here, we examine the relative effectiveness of different ways of learning new word meanings, finding that more ostensive experiences are not necessarily more effective than indirect learning via merely observing how a word is used.

Both research and intuition suggest that explicit/direct experiences with new words (often times via definitions or ostensive referents) are efficient and effective ways of learning new word meanings (Gruhn, et al., 2020; Watts, 1995). In comparison, the knowledge we gain from experience with words in natural text may seem somewhat fuzzy, imprecise, and variable from one instance to another. This variability, however, provides rich distributional information, helping link the new word to already known words.

One crucial aspect of word knowledge requires learners to generalize to new situations or different modalities. Though efficient, do these more explicit, direct experiences also yield generalizable word knowledge? Conversely, have we underestimated the richness that naturalistic text imparts during learning? In Experiment 1, we ask whether richer but less precise contexts (sentences), or more explicit/direct contexts (images and definitions) best yield generalization to other modalities or types of text. Experiment 2 builds on this finding by demonstrating that surprisingly little exposure is required for the distributional patterns of naturalistic text to efficiently impart word meaning.

Experiment 1

To test the how well different word learning experiences impart generalizable word knowledge, participants (N=58) were exposed to 12 novel word meanings (e.g. "the empty space at the top of a container") and pseudowords (Keuleers & Brysbaert, 2010) in one of three conditions where they either: read a definition, viewed four images depicting the new word's meaning, or read five sentences generated using ChatGPT (OpenAI, 2023) that used the word in context without defining it. To test how well participants learned the word meanings, we showed them new unlabeled images, definitions, and (cloze) sentences for each trained meaning and asked them to match it to one of the words presented. Because we were interested in generalization, our analysis only included responses for a given word if the participant answered correctly when tested in its exposure condition. A mixed effects logistic regression model was used to analyze the relationship between exposure condition and generalization X2 (1, N=58) = 4.79, p=.028 and participants who learned from sentences (M=.33, SE=.05) were more accurate in generalizing to other test conditions compared to participants who learned via images (M=.19, SE=.04) or definitions (M=.21, SE=.04). In sum, learning from passive exposure to text better supported generalization to situations that involved other types of word knowledge and visual knowledge.

Experiment 2

What do these results, then, say about human cognition? If we have underestimated the richness that linguistic experience affords during word learning, we may have also underestimated one of the processes believed to underlie word learning – distributional learning. To assess the role of distributional learning with minimal exposure, participants (*N*=86) learned three rare words (Brysbaert, et al., 2019) by reading ten sentence contexts sampled from COCA (Davies, 2008-). After exposure, participants provided definitions for the newly learned target words. A separate set of participants (N=30) defined and reported their familiarity with these words (without receiving any exposure). Sentence embeddings (Reimers & Gurevych, 2019) were then computed for the definitions collected from the experimental, high, and low familiarity groups, and evaluated for similarity to dictionary definitions. Bootstrapped means of these embedding similarities showed that participants with just ten exposures, *M*=.29, 95% CI [.28, .29], moved away from definitions of people who reported not knowing the word, *M*=.20, 95% CI [.19, .21], and towards definitions of those who reported knowing the word *M*=.37, 95% CI [.36, .38].

Our findings show how learners leverage the richness of natural language to gain generalizable, expert-like word meaning knowledge from surprisingly little exposure. Ongoing work is exploring how even more minimal text exposure and controlling for RTs may provide a window into relative efficiencies.

References

Brysbaert, M., Mandera, P., McCormick, S. F., & Keuleers, E. (2019). Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods*, *51*(2), 467–479. https://doi.org/10.3758/s13428-018-1077-9

Davies, Mark. (2008-) The Corpus of Contemporary American English (COCA). https://www.english-corpora.org/coca/.

Gruhn, S., Segers, E., & Verhoeven, L. (2020). Moderating role of reading comprehension in children's word learning with context versus pictures. *Journal of Computer Assisted Learning*, *36*(1), 29–45. https://doi.org/10.1111/jcal.12387

Hahn, E. R., & Gershkoff-Stowe, L. (2010). Children and Adults Learn Actions for Objects More Readily Than Labels. *Language Learning and Development*, *6*(4), 283–308. https://doi.org/10.1080/15475441003635315

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, *42*(3), 627–633. https://doi.org/10.3758/BRM.42.3.627

Nagy, W. E., Herman, P. A., & Anderson, R. C. (1985). Learning Words from Context. *Reading Research Quarterly*, *20*(2), 233–253. https://www.jstor.org/stable/747758.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 3982–3992. https://doi.org/10.18653/v1/d19-1410

Saragi, T., Nation, I. S. P., & Meister, G. F. (1978). Vocabulary learning and reading. *System*, *6*(2), 72–78. https://doi.org/10.1016/0346-251X(78)90027-1

Watts, S. M. (1995). Vocabulary instruction during reading lessons in six classrooms. *Journal of Literacy Research*, *27*(3), 399–424. https://doi.org/10.1080/10862969509547889

When a fishing rod becomes a tyre: on gesture comprehension in 2-to-3 years old transition in children

Monika Boruta-Żywiczyńska*[1], Marta Sibierska[1]

*Corresponding Author: monika.feba.boruta@gmail.com
[1]Center for Language Evolution Studies, Nicolaus Copernicus University in Toruń, Poland

Between the ages of 2 and 3, we observe a rapid growth in the spoken vocabulary of toddlers (Ganger & Brent 2004).However, the trajectory of co-speech or silent gesture production, as shown by Namy at al. (2004), is not characterised by constant growth. Language acquisition in the considered period can be characterised by a "trial-and-error" approach children adopt in everyday communication (Gentner & Namy, 2006; Benson, 2020). The trial-and error method used by toddlers is commonly described based on the mistakes they make in speech: (1) overgeneralisation, related to the use of improper syntactic structures (Baker, 1979; Onnis et al., 2002; Parke & Gauvain, 2009); (2) overextension, related to the use of a single word as a label for various objects (Rescorla, 1980; Clark, 2015; Barrett, 2017); and (3) underextension, when a child uses a word for a single item and does not see that the item belongs to a broader category (White, 1982; López- Couso et al., 2017; Barrett, 2017). In our presentation, we address the mistake of overextension in a gesture task. In the research, however, we did not look at speech. All of the errors occurred during a silent gesture comprehension.

The main experiment focused on children's ability to comprehend signs presented to them in the form of iconic gestures in three groups of children: 24-monthers, 30-monthers, and 36-monthers (total n=30). Each child was presented with a 36-pages-long book that contained 4 images per page. In 3 consecutive rounds, 12 pages per round, each child was asked to match the gestures of the experimenter with one image designated to the gesture. The task was challenging, because the children were shown two types of iconic gestures: enacting and representing ones. While statistical analysis revealed that children score higher with age, and that there is a change in preference for gesture comprehension: from representing gestures to enacting gestures, the qualitative viewing of the video material resulted in additional observations. In the post-experimental analysis, described here, we observed 66 examples of overextension mistake in gesture comprehension from 16 children. We observed that children made similar mistakes within their age groups---they extended one characteristic of a gesture presented to them by the experimenter (e.g. the spinning reel of the fishing rod in an enacting gesture) onto another---and in their answer pointed to the wheel present on the same page (ignoring the rest of the observed gesture indicating the fishing rod). The mistakes seem to be related to their experience with and knowledge about different kinds of objects and operations done to or with these objects.

In our analysis, we provide each pair (the expected and incorrect answer), describe the overextended manual characteristic of a given gesture, and try to the account for these mistakes using cognitive representation and prototype theory (Rosch 1975) from the perspective of gesture use. Intertwining these with Piaget's understanding of intellectual growth (adaptation and adjustment of knowledge), as well as the notion of schema (1952: 7; Inhelder & Piaget, 1958), we describe how the mistakes we observe provides us with insight into children's information processing in a comprehension task. Overextension is not only a mistake children make in speech---insights from silent gesture comprehension can help us understand how human thinking and conscious perception of characteristic features of actions and objects change and mature over time.

Acknowledgements

This work has been supported by the National Science Centre, Poland, under *Preludium* grant project no. UMO-2022/45/N/HS2/04155, entitled: "Let me show you! – Gestural repertoire of preschoolers in Poland."

References

Baker, C. L. (1979). Syntactic theory and the projection problem. *Linguistic Inquiry*, 10, 533–581.

Barrett, M. (2017). Early lexical development. *The handbook of child language*, 361–392.

Benson, J. B. (2020). *Encyclopedia of infant and early childhood development*. Elsevier.

Clark, E. V. (2015). Semantics and language acquisition. *The Handbook of Contemporary Semantic Theory*, 714–733.

Ganger, J., & Brent, M. R. (2004). Reexamining the vocabulary spurt. Developmental psychology, 40(4), 621.

Gentner, D., & Namy, L. L. (2006). Analogical processes in language learning. *Current Directions in Psychological Science*, 15(6), 297–301.

Inhelder, B., & Piaget, J. (1958). The growth of logical thinking from childhood to adolescence. New York: Basic Books.

López-Couso, M. J., Hundt, M., Mollin, S., & Pfenninger, S. E. (2017). Transferring insights from child language acquisition to diachronic change (and vice versa). *The changing English language: Psycholinguistic perspectives*, 332–347.

Namy, L. L., Campbell, A. L., & Tomasello, M. (2004). The changing role of iconicity in non-verbal symbol learning: A U-shaped trajectory in the acquisition of arbitrary gestures. Journal of Cognition and Development, 5(1), 37–57.

Onnis, L., Roberts, M., & Chater, N. (2002). Simplicity: A cure for overgeneralizations in language acquisition?. In *Proceedings of the Annual Meeting of the Cognitive Science Society* 24(24).

Parke, R. D. & Gauvain, M. (2009). *Child Psychology: A Contemporary Viewpoint*. New York, NY: McGraw-Hill.

Piaget, J. (1952). *Play, dreams and imitation in childhood*. W W Norton & Co.

Rescorla, L. A. (1980). Overextension in early language development. *Journal of child language*, *7*(2), 321–335.

Rosch, Eleanor (1975). "Cognitive representations of Semantic Categories". Journal of Experimental Psychology: General. 104 (3): 192–233.

White, T. G. (1982). Naming practices, typicality, and underextension in child language. *Journal of Experimental Child Psychology*, *33*(2), 324–346.

Past, Present, and Future: A literature review of the genetic research into the evolution of human language

Hannah Brown*[1]

*Corresponding Author: hannahbrown@cantab.net
[1]University of Victoria, Victoria, BC, Canada

1. Introduction

Since Forkhead box P2 (FOXP2) was first identified as being a gene involved in speech and language in 2001 (Lai et al., 2001) the amount of research into the genetic basis for human language has grown exponentially. Despite FOXP2 being initially hailed as the "language gene", over the last 20 years it has since become abundantly clear that the situation is much more complex than that (Deriziotis & Fisher, 2017; Fisher, 2019). Language is clearly a complex cognitive trait meaning that it is influenced by multiple genes and genomic pathways. To untangle this complex genetic architecture multiple streams of evidence (both genomic and otherwise) will need to be analysed together, including considering multiple genetic targets at once, to create a clearer picture (Deriziotis & Fisher, 2017; Eising et al., 2022).

2. Methods

A State-of-the-Art (SotA) literature review was conducted to analyse the current landscape of genetic research on the evolution of human language, alongside identifying historic trends in this research area, and themes for future research. This type of literature review specifically seeks to synthesise a summary of current thinking, examine how such perspectives may have changed over time, look at the historical trends within the research literature, and suggest areas for future directions of research (Barry et al., 2022b). This SotA review was conducted using Covidence (a cloud-based software) and the six stage methodological approach for SotA reviews suggested by Barry et al. (2022a). The process pathway for papers to be included in this review is shown in Fig. 1 below.

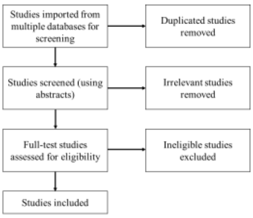


Figure 1. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram for this review

The Web of Science database was used for this review, using the search expressions “Language evolution genetic”, “Genome language evolution”, and “Gene human language”. From this search 9,585 studies were imported to Confidence for screening, 876 removed as duplicates, and 8,709 taken to manual screening (title and abstract, followed by full text review) and further analysis.

3. Discussion

3.1 Future research areas

One area suggested for future research is a shift away from “popular” genes to ensure that all genes of potential interest are equally investigated. For example, even in 2015 it was clear that some genes were more studied that others (Brown, 2015). This is likely as FOXP2 was dubbed “the language gene” upon its discovery, garnering substantial news interest and thus skewing research focus in this direction (Brown, 2015). As well as language being multigenetic, several transcription factors have been identified as involved in language evolution (Brown, 2019). Thus complex intersecting pathways are underlying language evolution, promoting a broad approach to this research. While still underrepresented in the literature there is beginning to be a shift towards this type of multigene/genome-wide work (Eising et al., 2022).

3.2 Limitations

A clear limitation of this review is that it was conducted with a single reviewer, which can compound the effects of the relativism and subjectivism that can be said to be inherent to this type of review (Barry et al., 2022a). When conducting a literature review we bring our own experiences, perspectives, and biases to the work, and thus this can affect the data interpretation. In order to conduct a stronger review it is suggested that this work be replicated with multiple reviewers, who may collaborate via Covidence to vote at each stage of the PRISMA process, with the goal of creating a consensus and reducing bias.

References

Barry, E. S., Merkebu, J., & Varpio, L. (2022a). *State-of-the-art literature review methodology: A six-step approach for knowledge synthesis* (5). *11*(5), Article 5. https://doi.org/10.1007/S40037-022-00725-9

Barry, E. S., Merkebu, J., & Varpio, L. (2022b). Understanding State-of-the-Art Literature Reviews. *Journal of Graduate Medical Education*, *14*(6), 659–662. https://doi.org/10.4300/JGME-D-22-00705.1

Brown, H.L. (2015) *The genetics of the evolution of human language: What can language deficits in modern humans tell us about our past?.* Undergraduate Dissertation. University of Cambridge, UK.

Brown, H.L. (2019) *What role have transcription factors played in the evolution of language?.* Masters Dissertation. University of Liverpool, UK.

Covidence systematic review software. (2023). [Computer software]. Veritas Health Innovation. https://www.covidence.org/

Deriziotis, P., & Fisher, S. E. (2017). Speech and Language: Translating the Genome. *Trends in Genetics*, *33*(9), 642–656.

Eising, E., Mirza-Schreiber, N., de Zeeuw, E. L., Wang, C. A., Truong, D. T., Allegrini, A. G., Shapland, C. Y., Zhu, G., Wigg, K. G., Gerritse, M. L., Molz, B., Alagöz, G., Gialluisi, A., Abbondanza, F., Rimfeld, K., van Donkelaar, M., Liao, Z., Jansen, P. R., Andlauer, T. F. M., … Fisher, S. E. (2022). Genome-wide analyses of individual differences in quantitatively assessed reading- and language-related skills in up to 34,000 people. *Proceedings of the National Academy of Sciences*, *119*(35), e2202764119. https://doi.org/10.1073/pnas.2202764119

Fisher, S. E. (2019). Human Genetics: The Evolving Story of FOXP2. *Current Biology : CB*, *29*(2), R65–R67. https://doi.org/10.1016/j.cub.2018.11.047 https://doi.org/10.1016/j.tig.2017.07.002

Lai, C. S. L., Fisher, S. E., Hurst, J. A., Vargha-Khadem, F., & Monaco, A. P. (2001). A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature*, *413*(6855), Article 6855. https://doi.org/10.1038/35097076

The cultural evolution of informative writing systems

Jon W. Carr*[1] and Kathleen Rastle[1]

*Corresponding Author: jon.carr@rhul.ac.uk
[1]Department of Psychology, Royal Holloway, University of London, England

The written and spoken forms of a language are subject to different evolutionary pressures. Over time, this can result in substantial divergence between the two, as each form of the language becomes better adapted to its own niche (Rastle, 2019). One example of this is the heterographic spelling of homophonous words, such as *knight* and *night*. Written wordforms such as these impose additional costs in learning but may be beneficial in reading because they reduce ambiguity. If the benefit in reading outweighs the cost in learning, heterography may be selected for in the evolution of writing systems. We investigate this possibility by experimentally simulating the evolution of orthographic systems using the iterated learning paradigm (Kirby, Tamariz, Cornish, & Smith, 2015), contrasting what happens in the presence and absence of communicative pressure for ambiguity avoidance.

We consider two possible mechanisms by which heterography might emerge (Berg & Aronoff, 2021). In Experiment 1, we consider *differentiation*, which involves the creation of new spellings or the repurposing of existing spellings to differentiate words that are homophonous in speech. For example, the words *plain* and *plane* were originally variant spellings of the same word, but they have taken on distinct meanings over time (Carney, 1994, p. 412). In Experiment 2, we consider the *conservation* mechanism, in which heterographic homophones arise as an epiphenomenon of sound change. For example, the words *meat* and *meet* are homophonous in modern English due to the /ɛː/–/eː/ merger that took place during the Great Vowel Shift, but their spellings are heterographic because they continue to reflect Middle English pronunciation (Wells, 1982, p. 140).

We created a simple 3$\times$3 stimulus space of colored shapes. The words for these stimuli consisted of a stem and a suffix, and participants were taught both the spelling and pronunciation. The stems—*buvi*-, *zeti*-, and *wopi*-, which represent shape—never changed over time, but the suffixes (explained below) *could* change. Participants were arranged in transmission chains, with each participant learning the orthographic output of the previous participant in the chain. We ran ten chains of nine generations in each of two conditions: Transmission-only, in which participants were simply tested on the orthographic system they had been trained on, and Transmission + Communication, in which each generation consisted of a pair

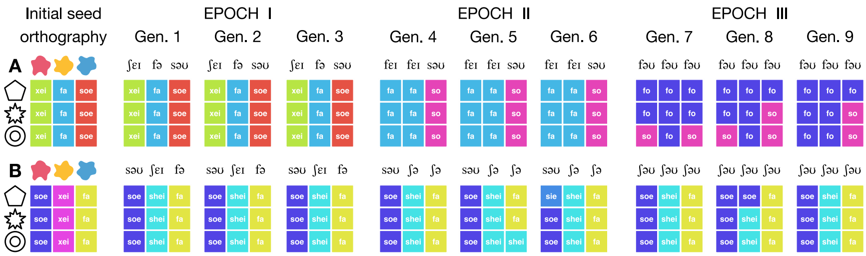


Figure 1. Suffix spellings in two example chains from Experiment 2. Each color represents a unique suffix spelling. **A** Transmission-only condition. The orthography transparently reflects the increasing homophony but, as a result, becomes unable to express the color dimension. **B** Transmission + Communication condition. The orthography is conserved in the face of increasing homophony, allowing the system to express color at the expense of transparency.

of participants who played a communication game that incentivized ambiguity minimization (following similar methods to Kirby et al., 2015).

In Experiment 1, which tests the differentiation mechanism, the suffixes were always pronounced /-kəʊ/, but the orthography was seeded with high variation, such that the suffix could be spelled in many different ways using the graphemes ⟨c⟩, ⟨k⟩, ⟨q⟩, ⟨o⟩, ⟨oe⟩, and ⟨oh⟩. We hypothesized that, under communicative pressure, the orthographies would be more likely to adopt differentiated suffix spellings conditioned on color (e.g., ⟨-co⟩, ⟨-koh⟩, and ⟨-qoe⟩ for pink, yellow, and blue), despite all colors being expressed homophonously in speech (i.e., /-kəʊ/). However, the results revealed little evidence of differentiation. In most cases, the orthographic systems simply became transparent—a single spelling was adopted for the suffix regardless of color, even under communicative pressure.

In Experiment 2, which tests the conservation mechanism, the initial seed systems were entirely regular and compositional, with distinct suffixes for each color (e.g., /-səʊ/, /-fə/, and /-ʃɛɪ/ spelled ⟨-soe⟩, ⟨-fa⟩, and ⟨-xei⟩). Over three epochs, we experimentally induced sound changes that resulted in increasing homophony. We hypothesized that, under communicative pressure, the orthography would be more likely to remain intact, continuing to express color at the cost of transparently mirroring the homophony. Indeed, this is what we observed across several chains; an example is shown in Fig. 1.

Our findings suggest firstly that pressure for informativeness (induced through communicative pressure) can give rise to spellings that are more expressive than their spoken counterparts, and secondly that informative heterography is easier to attain through the conservation (as opposed to differentiation) mechanism. We further discuss how these small-scale simulations can inform our understanding of the real-world processes underlying spelling change, including the roles of variation, redundancy, top-down reform, and other functional explanations.

Acknowledgments

This work was funded by a Leverhulme Trust Research Project Grant (grant number: RPG-2020-034).

References

Berg, K., & Aronoff, M. (2021). Is the English writing system phonographic or lexical/morphological? A new look at the spelling of stems. *Morphology*, *31*(3), 315–328.

Carney, E. (1994). *A survey of English spelling.* Routledge.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102.

Rastle, K. (2019). EPS mid-career prize lecture 2017: Writing systems, reading, and language. *Quarterly Journal of Experimental Psychology*, *72*(4), 677–692.

Wells, J. C. (1982). *Accents of English* (Vol. 1). Cambridge University Press.

Which came first—iconicity or symbolism?

Erica A. Cartmill*[1], Matt Cartmill[2], Kaye Brown[2], and Jacob G. Foster[1]

*Corresponding Author: ericac@iu.edu
[1]Cognitive Science, Indiana University, Bloomington, IN, USA
[2]Evolutionary Anthropology, Duke University, Durham, NC, USA

Iconicity has long played a privileged role in theories of language origins, purportedly helping to “cold start” language by making meanings more transparent. Theories proposing a gestural origin of language hinge on gesture’s ability to illustrate things iconically (Armstrong et al. 1995, Corballis 2002, Zlatev 2008; Arbib 2012). So too do theories proposing that the earliest words were onomatopoetic (Fitch 2010). Iconicity provides an attractive solution to the problem of how to get language started. If a human ancestor invented a word for something, how would anyone know what it meant, especially if they didn’t have other words to explain it? Iconicity solves this problem by proposing that meanings could be acted out, and that similarities between icon and referent provide insight into meaning. This is a seductive proposal, but it assumes that early iconic reference was (1) easy to produce and (2) easy to understand. There is some evidence that iconicity may function this way in adult humans, at least with contextual pragmatic support, as in experiments where someone guesses the meaning of a pantomime from a set of alternatives (e.g., Sibierska et al. 2022). But iconic reference is not easily understood by either young children or non-human apes—both of which have an easier time with conventional symbols. Importantly, symbol-like markings also precede iconic drawings in the archaeological record. All these sources suggest that symbolism (achievable via associative learning) preceded iconic reference in the evolution of language.

Although children gesture before they can speak (Bates et al. 1979), iconic gestures are rare and develop later than conventional gestures (Özçalişkan & Goldin-Meadow 2011). This is true even in homesign (manual systems created by deaf children with no access to language models), despite iconicity's being much more prevalent in homesign (Cartmill et al. 2017).

Like children, great apes are prolific gesturers, but iconicity is almost entirely absent from their natural communication (Call & Tomasello 2007). There are a few examples of pantomime-like gestures in language-trained or rehabilitant apes (Russon & Andrews 2011; Perlman & Gibbs 2013), but it is difficult to rule out the possibility that apes copy human movements without understanding the iconic mappings beneath them. Understanding icon-to-world mappings is not trivial. Judy DeLoache argues that in order to do this, children must represent an object simultaneously as both an object and as a representation of another object (DeLoache 1995). Studies suggest that children do not begin to master this ability until the ages of 3-4. Majid and Pyers (2017) found that children were not able to guess the meanings of iconic gestures until 4-5 years-old. Even children learning sign language can struggle with iconicity. Signing children do not master classifiers (which rely on iconic mapping) until 5-9 (Mayberry & Squires 2006).

Symbolic signs also precede iconic representation in another visual medium: drawing. Representational (iconic) art first appears in the archaeological record in connection with anatomically modern humans, about 45,000 years ago (Brumm et al. 2021). However, purely abstract symbolic markings were made much earlier by both Neandertals in Europe and early modern humans in Africa, at least 65,000 and 75,000 years ago respectively (Henshilwood et al. 2002; Hoffmann et al. 2018; Garcia-Diez 2022). Similar abstract "drawings" were made around 500,000 years ago by *Homo erectus* on Java (Joordens et al. 2015). We conclude from this that symbolic conventions surrounding the making of marks on surfaces were in place long before markings were used to represent iconically.

Great apes have little difficulty learning to use conventional referential symbols like those of Yerkish and ASL, but have considerable difficulty understanding representational drawings (Close & Call 2015; Martinet & Pelé 2021). One study found that apes and children under 3 were successful at finding hidden rewards when they were labeled with arbitrary symbols, but not when they were marked with iconic drawings (Tomasello et al. 1997). Children also struggle with 3D iconic representations. Children under 3 struggle to find hidden rewards when shown the location using a model of the room, but succeed when given verbal (symbolic, conventional) instructions (DeLoache & Burns 1993).

These very different lines of evidence—from studies of child development, from experiments on great apes, and from the archeological record—point to the same conclusion: the use and interpretation of iconic signs involves sophisticated cognitive abilities that appear relatively late, both in human development and in human evolution. All this evidence suggests that symbolic reference, not iconic representation, provided the framework for the earliest steps towards language.

References

Arbib, M.A. (2005). From monkey-like action recognition to human language: An evolutionary framework for neurolinguistics. *Behavioral and Brain Sciences, 28*, 105–167.

Armstrong, D.F., Stokoe, W.C., Wilcox, S.E. (1995). *Gesture and the Nature of Language*. Cambridge, UK: Cambridge Univ. Press.

Bates, E., & Dick, F. (2002). Language, gesture, and the developing brain. *Developmental Psychobiology*, *40*(3), 293-310.

Brumm, A., Oktaviana, A.A., Burhan, B., Hakim, B., Lebe, R., Zhao, J.X., Sulistyarto, P.H., Ririmasse, M., Adhityatama, S., Sumantri, I. and Aubert, M., (2021). Oldest cave art found in Sulawesi. *Science Advances*, *7*(3), eabd4648.

Call, J., & Tomasello, M. (2007). Comparing the gestural repertoire of apes. In J. Call & M. Tomasello (Eds.), *The Gestural Communication of Apes and Monkeys* (pp. 197–220). Mahwah, NJ: Erlbaum.

Cartmill, E. A., Rissman, L., Novack, M. A., & Goldin-Meadow, S. (2017). The development of iconicity in children's co-speech gesture and homesign. *Language, Interaction and Acquisition*, *8*(1), 42-68.

Close, J., & Call, J. (2015). From colour photographs to black-and-white line drawings: an assessment of chimpanzees' (*Pan troglodytes'*) transfer behaviour. *Animal Cognition*, *18*, 437-449.

DeLoache, J. (1995). Early understanding and use of symbols: The model model. *Current Directions in Psychological Science, 4*, 109–113.

DeLoache, J., & Burns, N.M. (1993). Symbolic development in young children: Understanding models and pictures. In C. Pratt & A. F. Garton (Eds.), *Systems of Representation in Children*. London: John Wiley & Sons.

Fitch, W. T. (2010). *The Evolution of Language.* Cambridge, UK: Cambridge University Press.

Garcia-Diez, M. (2022). 'Art': Neanderthal symbolic graphic behaviour. In F. Romagnioli, F. Rivals and S. Benazzi (Eds.), *Updating Neanderthals: Understanding Behavioral Complexity in the Late Middle Paleolithic* (pp. 251-260). London: Academic Press.

Henshilwood, C., d'Errico, F., Yates, R., Jacobs, Z., Tribolo, C., Duller, G.A.T., Mercier, N., Sealy, J.C., Valladas, H., Watts, I., and Wintle, A.G. (2002). Emergence of modern human behavior: Middle Stone Age engravings from South Africa. *Science 295*: 1278-1280.

Hoffmann, D. L. , Standish C.D., García-Diez, M. Pettitt, P.B. Milton, J.A., Zilhão, J., Alcolea-González, J.J., Cantalejo-Duarte, P., Collado, H., de Balbín, R., Lorblanchet, M., Ramos-Muñoz,, J., Weniger, G.- Ch., and Pike, A. W. G. (2018). U-Th dating of carbonate crusts reveals Neandertal origin of Iberian cave art. *Science 359*: 912-915.

Iverson, J.M., Fagan, M.K. (2004). Infant vocal–motor coordination: Precursor to the gesture–speech System? *Child Development 75*: 1053–1066.

Joordens, J.C.A., d'Errico, F., Wesselingh, F.P., Munro, S., de Vos, J., Wallinga, J., Ankjærgaard, C., Reimann, T., Wijbrans, J.R., Kuiper, K.F., Mücher, H.J., Coqueugniot, H., Prie, V., Joosten, I., van Os, B., Schulp, A.S., Panuel, M., van der Haas, V., Lustenhouwer, W., Reijmer, J.J.G., and Roebroeks, W. (2015). *Homo erectus* at Trinil on Java used shells for tool production and engraving. *Nature, 518*: 228-231.

Magid, R.W., & Pyers, J. E. (2017). "I use it when I see it": The role of development and experience in Deaf and hearing children's understanding of iconic gesture. *Cognition*, *162*, 73-86.

Martinet, L., & Pelé, M. (2021). Drawing in nonhuman primates: What we know and what remains to be investigated. *Journal of Comparative Psychology*, *135*(2), 176.

Mayberry, R.I., & Squires, B. (2006). Sign language acquisition. *Encyclopedia of language and linguistics*, *11*, 739-43.

Özçalişkan, Ş., & Goldin-Meadow, S. (2011). Is there an iconic gesture spurt at 26 months? In G. Stam and M. Ishino (Eds.), *Integrating gestures: The interdisciplinary nature of gesture.* Amsterdam, NL: John Benjamins.

Perlman, M., & Gibbs, R. W. (2013). Pantomimic gestures reveal the sensorimotor imagery of a human-fostered gorilla. *Journal of Mental Imagery*, *37*(3/4), 73-96.

Russon, A.E., & Andrews, K. (2011a). Orangutan pantomime: elaborating the message. *Biology Letters*, *7*(4), 627-630.

Sibierska, M., Boruta-Żywiczyńska, M., Żywiczyński, P., & Wacewicz, S. (2022). What's in a mime? An exploratory analysis of predictors of communicative success of pantomime. *Interaction Studies*, *23*(2), 289-321.

Tomasello, M., Call, J., & Gluckman, A. (1997). Comprehension of novel communicative signs by apes and human children. *Child Development*, 1067-1080.

Zlatev, J. (2008). The co-evolution of intersubjectivity and bodily mimesis. *The Shared Mind: Perspectives on Intersubjectivity*, 215-244.

From speech to primate vocalizations: Self-supervised deep learning as a comparative approach

Jules Cauzinille[*1,2,3], Benoît Favre[1,2], Ricard Marxer[1,2], and Arnaud Rey[1,3]

[*]Corresponding Author: jules.cauzinille@lis-lab.fr
[1]ILCB, Aix-Marseille University, Marseille, France
[2]LIS, Aix-Marseille University, Marseille, France
[3]CRPN, Aix-Marseille University, Marseille, France

Within the recent deep learning revolution, *transformer* architectures and *pre-trained self-supervised models* opened up many perspectives for the study of linguistics and animal communication. These state-of-the-art tools efficiently address a wide range of applications in monitoring animal behavior through sound (Stowell, 2022; Kahl, Wood, Eibl, & Klinck, 2021; Hagiwara, 2023) or in assisting humans with language related tasks. The increasing scientific interest generated by this revolution raises the following question: can acoustic deep learning be leveraged as a scientific tool in the study of the evolution of language?

We propose a novel methodology involving the use of deep learning models as comparative toolkits by testing their ability to jointly process speech and non-human vocal communication. This approach relies on the disentanglement of self-supervised learning (SSL) pre-trained models, i.e., computer models trained on large unlabeled datasets. SSL models were introduced in the field of computer vision (Jing & Tian, 2021) as a way to leverage the extensive availability of image data. They rely on the assumption that pre-training a first model to encode and extract information from large collections of raw data can benefit secondary models specialized in downstream tasks on smaller-sized datasets. By applying this method to acoustic data, researchers were able to develop efficient speech processing models, outperforming most purely *supervised* solutions (Mohamed et al., 2022; Yang et al., 2021). SSL models trained on speech datasets show high performance on an array of tasks (Evain et al., 2021) and learn to encode different levels of linguistic information during pre-training without the need for supervision. For instance, Pasad, Shi, and Livescu (2023) showed that low-level acoustic information tends to be encoded in the initial layers of these models while high-level phonemic or lexical information is mostly encoded in deeper layers.

By adapting the SSL approach to bioacoustic tasks, we develop transfer learning experiments aimed at understanding how much information speech-based models are able to extract from non-human vocalizations. We focus our preliminary experiments on non-human primates, more specifically apes, as our closest living relatives provide a unique opportunity to explore the evolutionary basis of our vocal communicative behavior. We rely on models pre-trained on human speech (Hsu et al., 2021; Schneider, Baevski, Collobert, & Auli, 2019) to perform primate-related bioacoustic tasks and compare them to models pre-trained on other taxa such as birds (Kahl et al., 2021), or general acoustic data such as music, video soundtracks, etc. (Huang et al., 2022; Kong et al., 2020). The tasks include vocal identity recognition, detection of vocalizations in natural contexts and call-type classification.

We define three main approaches to test the knowledge transfer capabilities of SSL models from speech to primate vocalizations. The **probing** approach consists in using pre-trained models as feature extractors. Said features are then "probed" with logistic regression to disentangle the type of information they extracted from primate vocalizations. Good performance on a given task shows that the information needed to answer the task was successfully extracted during pre-training and is linearly separable within the model's representations. The **fine-tuning** approach involves further training SSL models on small datasets to improve their performance on the downstream task. It can show how much more training data a model needs to efficiently extract information from primate vocalizations. Finally, to ensure true knowledge transfer from human to other primates, a third method involves **parameter-efficient fine-tuning** (PEFT) and **adversarial reprogramming** (Elsayed, Goodfellow, & Sohl-Dickstein, 2018; Zheng et al., 2023). Both methods allow keeping the pre-trained weights of the original model untouched by training additional "filters" for primate-related tasks.

Preliminary experiments consist in recognizing vocal signatures of individual gibbons (*Hylobates funereus*). The probing method shows that the initial layers of speech-based models are capable of extracting sufficient information to classify the individual voices of 10 female gibbons with up to 95% accuracy. This result outperforms models pre-trained on birdsongs, which seem to heavily rely on recognizing the background noise of recordings rather than the primate's vocal signature. Additionally, we demonstrate the ability of some speech models to recognize gibbon's vocal identities from the temporal dynamics of their song rather than the anatomical specificities of their voices. Finally, when the fine-tuning method is applied, further performance gains can be observed, even in few-shot learning setups.

This type of result helps us examine divergences and similarities between speech and primate vocalizations from a deep learning perspective. They show how speech-based pre-training may be at an advantage when dealing with primate vocal communication by transferring knowledge from one to the other. In general terms, our experiments test for the validity of deep transfer learning as a scientific tool in the study of the origins of language from a comparative standpoint. Future experiments will focus on extending previously mentioned methods to other tasks and primate species.

Acknowledgments

This work, carried out within the Institute of Convergence ILCB (ANR-16-CONV-0002), has benefited from support from the French government (France 2030), managed by the French National Agency for Research (ANR) and the Excellence Initiative of Aix-Marseille University (A*MIDEX). It was also supported by the COMPO ANR project (#ANR-23-CE23-0031) and the HEBBIAN ANR project (#ANR-23-CE28-0008).

All gibbon recordings and annotations were provided by Dena J. Clink (K. Lisa Yang Center for Conservation Bioacoustics and Cornell Lab of Ornithology, Cornell University, Ithaca, NY, USA). More information can be found in Clink, Kier, Ahmad, and Klinck (2023) and Clink, Bernard, Crofoot, and Marshall (2017)

References

Clink, D., Kier, I., Ahmad, A., & Klinck, H. (2023). A workflow for the automated detection and classification of female gibbon calls from long-term acoustic recordings. *Frontiers in Ecology and Evolution*, *11*.

Clink, D. J., Bernard, H., Crofoot, M. C., & Marshall, A. J. (2017). Investigating Individual Vocal Signatures and Small-Scale Patterns of Geographic Variation in Female Bornean Gibbon (Hylobates muelleri) Great Calls. *International Journal of Primatology*, *38*(4), 656–671.

Elsayed, G. F., Goodfellow, I., & Sohl-Dickstein, J. (2018). *Adversarial Reprogramming of Neural Networks.* arXiv. (arXiv:1806.11146 [cs, stat])

Evain, S., Nguyen, H., Le, H., Zanon Boito, M., Mdhaffar, S., Alisamir, S., Tong, Z., Tomashenko, N., Dinarelli, M., Parcollet, T., Allauzen, A., Estève, Y., Lecouteux, B., Portet, F., Rossato, S., Ringeval, F., Schwab, D., & Besacier, L. (2021). LeBenchmark: A Reproducible Framework for Assessing Self-Supervised Representation Learning from Speech. In *INTERSPEECH 2021: Conference of the International Speech Communication Association.* Brno, Czech Republic.

Hagiwara, M. (2023). Aves: Animal vocalization encoder based on self-supervision. In *Icassp 2023 - 2023 ieee international conference on acoustics, speech and signal processing (icassp)* (p. 1-5).

Hsu, W.-N., Bolte, B., Tsai, Y.-H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *PP*, 1-1.

Huang, P.-Y., Xu, H., Li, J., Baevski, A., Auli, M., Galuba, W., Metze, F., & Feichtenhofer, C. (2022). Masked autoencoders that listen. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in neural information processing systems* (Vol. 35, pp. 28708–28720). Cur-

ran Associates, Inc.
Jing, L., & Tian, Y. (2021). Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(11), 4037-4058.
Kahl, S., Wood, C. M., Eibl, M., & Klinck, H. (2021). Birdnet: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, *61*, 101236.
Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., & Plumbley, M. D. (2020). Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 2880-2894.
Mohamed, A., Lee, H. yi, Borgholt, L., Havtorn, J., Edin, J., Igel, C., Kirchhoff, K., Li, S.-W., Livescu, K., Maaløe, L., Sainath, T., & Watanabe, S. (2022). Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing*, *16*(6), 1179–1210.
Pasad, A., Shi, B., & Livescu, K. (2023). Comparative layer-wise analysis of self-supervised speech models. In *Icassp 2023-2023 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 1–5).
Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). wav2vec: Unsupervised Pre-Training for Speech Recognition. In *Proc. interspeech 2019* (pp. 3465–3469).
Stowell, D. (2022). Computational bioacoustics with deep learning: a review and roadmap. *PeerJ*, *10*.
Yang, S., Chi, P., Chuang, Y., Lai, C., Lakhotia, K., Lin, Y., Liu, A., Shi, J., Chang, X., Lin, G., Huang, T., Tseng, W., Lee, K., Liu, D., Huang, Z., Dong, S., Li, S., Watanabe, S., Mohamed, A., & Lee, H. (2021). Superb: Speech processing universal performance benchmark. In *Icassp 2023-2023 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 1–5). International Speech Communication Association.
Zheng, Y., Feng, X., Xia, Z., Jiang, X., Demontis, A., Pintor, M., Biggio, B., & Roli, F. (2023). Why adversarial reprogramming works, when it fails, and how to tell the difference. *Information Sciences*, *632*, 130–143.

Discreteness and systematicity emerge to facilitate communication in a continuous signal-meaning space

Alicia M Chen[*1], Matthias Hofer[*1], Moshe Poliak[1], Roger Levy[1], and Noga Zaslavsky[2]

[1]Department of Brain and Cognitive Sciences, MIT
[2]Department of Language Science, UC Irvine
[*]Equal contribution, correspondence to aliciach@mit.edu

Language relies on the interplay of many intricate features to ensure that the richness and complexity of human experience can be communicated in a tractable way. Two of these features are *discreteness* and *systematicity*. Discreteness provides a segmentation of inherently continuous phonetic and semantic spaces into distinguishable units and categories, while systematicity allows for these elements to be aligned in organized ways, ensuring that language is not only highly efficient but also predictably expressive.

Previous research has explored the emergence of these properties independently, highlighting the role of systematicity in language acquisition (Dingemanse, Blasi, Lupyan, Christiansen, & Monaghan, 2015), use (Nölle, Staib, Fusaroli, & Tylén, 2018), and its transmissibility and evolvability (Kirby, Cornish, & Smith, 2008). Conversely, work on discreteness has focused on its emergence in continuous signaling spaces along with combinatoriality (Verhoef, 2012; Little, Eryılmaz, & De Boer, 2017).

However, the question of how systematicity and discreteness arise jointly to support efficient communication — especially when *both* the signal and meaning spaces are continuous — and how these properties might constrain or reinforce one another, has been largely unexplored. In this study, we examine the concurrent emergence of these features in a two-player communication experiment where participants were asked to generalize learned continuous signals to communicate about a continuous color space. The signal space was whistled signals produced by an on-screen slide whistle interface, and the meaning space was defined by a subset of colors from the World Color Survey's standard color naming grid.

The experiment consisted of a learning phase and a communication phase. Participants learned 5 signal-color mappings. Five signals with a diverse set of perceptual features were selected from a larger set of signals collected by Hofer and Levy (2019), and their corresponding color referents were randomly selected to be approximately evenly spaced in hue. After learning, participants were paired up and asked to generalize those mappings to a larger set of color chips in a refer-

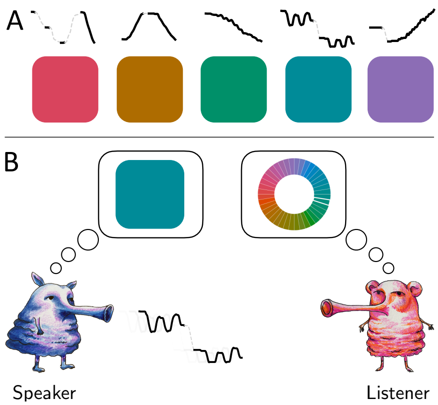


Figure 1. Our experimental framework. **A.** Initial color-signal pairings used in the learning phase. **B.** In the communication phase, speakers are presented with a target color and communicates it to the listener by generating a whistled signal. The listener guesses by selecting one of the 40 colors of the color wheel.

ence game. Our primary interest was in whether participants' extrapolated signals displayed elements of discreteness or systematicity in ways that supported successful communication. Discreteness was measured by calculating the cluster tendency of participants' signaling repertoires using the Hopkins statistic (Banerjee & Dave, 2004). Systematicity was measured by the correlation between pairwise signal distances (as measured using Dynamic Time Warping) and pairwise color distances in perceptually uniform CIELUV space (Schanda, 2007).

We found that participants learned to communicate successfully and aligned their signal repertoires, with more successful dyads showing higher degrees of alignment, suggesting that the formation of communicative conventions was crucial in driving communication performance. Furthermore, we observed the emergence of both systematicity and discreteness. However, we found that systematicity, but not discreteness, was correlated with better communication. Additionally, we note cases where participants seemed to have created composite signals to generalize to unseen colors, inviting speculation about the role of combinatoriality in this domain.

A few limitations of the present study include issues related to small-scale initializations in signal-meaning repertoires and limitations in measuring signal structure and similarity. Possible future extensions of this work are outlined, including investigating the role of discreteness and extending this setup to a multi-generational transmission experiment. Ultimately, we believe that these results contribute to a larger body of work exploring the role of human cognitive biases toward structure in the development and emergence of communication systems.

References

Banerjee, A., & Dave, R. N. (2004). Validating clusters using the hopkins statistic. In *2004 ieee international conference on fuzzy systems (ieee cat. no. 04ch37542)* (Vol. 1, pp. 149–153).

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, Iconicity, and Systematicity in Language. *Trends in Cognitive Sciences*, *19*(10), 603–615.

Hofer, M., & Levy, R. P. (2019). *Iconicity and Structure in the Emergence of Combinatoriality* (Preprint). PsyArXiv.

Kirby, S., Cornish, H., & 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.

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

Nölle, J., Staib, M., Fusaroli, R., & Tylén, K. (2018). The emergence of systematicity: How environmental and communicative factors shape a novel communication system. *Cognition*, *181*, 93–104.

Schanda, J. (2007). *Colorimetry: understanding the cie system.* John Wiley & Sons.

Verhoef, T. (2012). The origins of duality of patterning in artificial whistled languages. *Language and cognition*, *4*(4), 357–380.

Languages of esoteric societies provide a window into a previous stage in the evolution of human languages

Sihan Chen[1,], David Gil[2] and Antonio Benítez-Burraco[*3],

[*]Corresponding Author: abenitez8@us.es
[1]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA
[2]Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
[3]Department of Spanish, Linguistics & Theory of Literature, University of Seville, Spain

Contemporary languages of today's esoteric societies are argued to provide a model for a previous stage in the evolution of human languages. A recent analysis of data from the World Atlas of Language Structures compares languages spoken by esoteric and exoteric societies, showing that the languages of esoteric societies tend to be associated with more complex morphological structures alongside greater simplicity in the realm of syntax. Such correlations between societal and linguistic features provide a window into linguistic phylogeny. Given that until recently all human societies were highly esoteric, it may be inferred that the languages spoken by such Upper Paleolithic societies were similar to those of current esoteric societies, instantiating a earlier stage in the evolution of language characterized by more complex morphology but simpler syntax.

1. Introduction

There is a growing consensus that the evolution of human language and human linguistic capabilities was gradual rather than abrupt (Progovac, 2019), this paralleling the gradual evolution of the human physical and behavioral distinctive phenotype (Neubauer et al., 2018; Scerri and Will, 2023). This in turn raises the question what such intermediate stages in language phylogeny might have looked like. Addressing this question, a number of proposals have been put forward in an attempt to characterize early stages in the evolution of language. In one of the earliest and most renowned of such proposals, Bickerton (1990) posits a *protolanguage*, endowed with just rudimentary mechanisms for juxtaposing simple words together. Somewhat further down the evolutionary line, Gil (2017) posits an *IMA language* that is Isolating (lacking internal word structure), Monocategorial (lacking distinct parts of speech), and Associational (lacking construction-specific rules of semantic compositionality) — though with a more developed combinatorial syntax, and bearing a closer resemblance to at least some contemporary human languages, such as Riau Indonesian. Even later in the evolutionary trajectory, Benítez-Burraco and Progovac (2020) propose a language type characterized by substantially increased morphological complexity alongside a syntax still lacking some of the functional categories of many modern languages (e.g. complementizers), arguing that such languages were associated with the esoteric, inward-oriented societies typical of past stages of human evolution, and to a lesser extent also some contemporary societies such as those of hunter gatherers.

This paper provides novel empirical support for the latter proposal by Benítez-Burraco and Progovac of a stage in the evolution of language associated with Upper Paleolithic societies and characterized by rich morphology and relatively simple syntax. Our argument consists of two parts. First, we present the results of a recent large-scale survey of contemporary languages (Chen et al., 2023), demonstrating that societal esotericity correlates positively with morphological complexity, but negatively with syntactic complexity. Then, invoking the evolutionary inference principle for linguistic and cultural/socio-political complexity (Gil, 2021), we suggest that the languages of today's esoteric societies syntax, provide a model for the languages spoken by similarly esoteric societies in the evolutionary past.

2. Languages of esoteric and exoteric societies

Recent studies have examined the potential relationships between linguistic and societal structures. While some studies, e.g. Koplenig (2019) and Shcherbakova et al (2023) have not found evidence for such connections, a wide range of other studies have revealed some of the ways in which contemporary human languages spoken by esoteric societies differ systematically from their counterparts spoken by exoteric societies. Many of these studies make reference to the notion of complexity, both in the linguistic and societal domains. Specifically, exoteric societies have been characterized as more politically complex than their esoteric counterparts.

Several studies have demonstrated negative correlations between aspects of societal and linguistic complexity. As argued by McWhorter (2005, 2011, 2018), Dahl (2004), Wray and Grace (2007), Lupyan and Dale (2010), Trudgill (2011) and others, smaller societies, generally characterized by sociopolitical esotericity and more context-dependent forms of communication, are fertile grounds for the accretion of linguistic complexity in the domain of morphology, while larger political entities, typically associated with reduced sociopolitical esotericity and various modes of less context-dependent communication, particularly those able to convey propositional content to strangers, are conducive to linguistic simplification, specifically, in the domain of morphology, one possible reason being imperfect adult second-language acquisition.

In contrast to the above, however, a range of other studies support an opposite positive correlation between sociopolitical complexity and various aspects of linguistic complexity. Thus, recent experiments by Raviv, Meyer and Lev-Ari (2019, 2020) and Raviv (2020) show that in artificial languages, larger speech communities create more highly compositional languages — which entails increased complexity in the domain of combinatorial syntax. Similarly, in sign languages, Meir et al (2012) and Ergin et al (2020) argue that an increase in the size of the signing community results in a greater degree of conventionalization. In the realm of metaphor comprehension, Gil and Shen (2021) show that more highly complex polities tend to be associated with languages whose metaphors are endowed with more complex directional structure. With regard to Tense-Aspect-Mood marking, Gil (2021) demonstrates that languages belonging to larger families, the product of demographic spread, are associated with more complex systems characterized by obligatory as opposed to optional marking. Finally, in the domain of basic clause structure, work reported on in Gil and Shen (2019) shows that more highly complex polities tend to be associated with languages endowed with a greater degree of grammaticalization of thematic-role assignment.

How might these seemingly conflicting results be reconciled? The key lies in the observation that the linguistic features whose complexity correlates with societal complexity in opposing ways, either negatively or positively, are of two qualitatively different kinds. Simplifying somewhat, negative correlations between societal and linguistic complexity are characteristic of features of a morphological nature, while positive correlations between societal and linguistic complexity are associated with features of a syntactic nature.

We have found evidence to this effect in our research (preliminarily described in Benítez-Burraco et al., 2022; Chen et al., 2023). We classified the 82 out of 142 language features from the *World Atlas of Language Structures*, or *WALS* (Haspelmath et al., 2005) that are related to morphology or syntax, as purely morphological features (M), purely syntactic features (S), features pertaining to both domains but predominantly related to morphology (Ms) and features pertaining to both domains but predominantly related to syntax (mS). Independently, we characterized the diverse values for each feature in terms of complexity as either equipollent or privative: while equipollent features are ones in which there is no prima facie reason to characterize one of the feature values as more complex than the other, privative features are those in which different feature values may be ranked along a scale of complexity, with some feature values more complex than others. The analysis invokes the notion of descriptive complexity, considering one feature value to be more complex than another if its description makes use of a larger number of symbols. For example, WALS feature 22, "Inflectional Synthesis of the Verb", is first classified as primarily morphological (Ms), since it pertains to changes in word form, even though these different forms may be used secondarily for syntactic purposes, as in agreement. Having a larger number of inflectional forms is then taken to be indicative of greater morphological complexity. Likewise, WALS feature 81, "Order of Subject, Object and Verb", is first classified as purely syntactic (S). Then, free word order languages are regarded as being of lesser syntactic complexity than languages with a single dominant order.[1]

As for societal complexity, languages are ranked in accordance with a range of criteria drawn from three different sources: the Expanded Graded Intergenerational Disruption Scale (EGIDS), from *Ethnologue* (Eberhard et al., 2022); the size of the family to which the language belongs, from *Glottolog* (Hammarström et al., 2022); and a variety of criteria from the *D-Place* database (Kirby et al., 2016), including the number of jurisdictional levels above the local community (Feature EA033 in the database), the size of local communities (EA031), and population size (EA202).

Bringing together the above sources, we constructed a dataset containing 94 different classifications along with 1 societal PC. We ran a linear regression between each combination of a classification and the PC, resulting in 94 statistical tests. For binary classifications, namely those with only two values, we ran a logistics regression instead. For each statistical test, we reported the estimated slope along with the p-value. We say a relation between a principal component is significant if the p-value is less than 0.05. We also controlled for potential confounding factors, particularly, language family and geographical regions, by conducting an additional analysis in which we considered the phylogeny and the geographical proximity of languages.

Our results reveal a statistically significant tendency for simpler esoteric societies to be associated with languages of greater morphological complexity but lesser syntactic complexity than their more complex exoteric counterparts. Based

[1] In addition, WALS feature 81 distinguishes between six dominant word orders; however, since there is no obvious basis for characterizing one such order as more complex than another, this further distinction is considered to be equipollent and therefore ignored in the present analysis.

on these results, two language types are defined: *S-languages*, associated with eSoteric societies, exhibiting simpler syntax but more complex morphology, and *X-languages*, associated with eXoteric societies, characterized by more complex syntax but simpler morphology. Although esotericity and exotericity constitute two poles on a single scale of sociopolitical complexity, the factors driving the development of Type S and Type X languages are not mirror-images but rather of diverse and qualitatively different natures. Thus, while the correlation between esotericity and morphological complexity could be due to factors such as simplification resulting from imperfect adult second-language acquisition, the correlation between exotericity and syntactic complexification may be attributed to factors such as the need to satisfy a broader range of communicative needs, e.g. conveying more complex meanings to unrelated people. Moreover, since these two language types are based on quantitative analyses, they are most appropriately considered to be prototypes around which languages tend to cluster. In particular, as noted, many of the WALS features are of a mixed morphological/syntactic nature (Ms or mS in our characterization). For such features, then, the factors driving the development of Type S and Type X languages pull in opposite directions. For this reason, the development of Type S and Type X languages does not necessarily result, as might have been expected, in a strict trade-off between the morphological and syntactic complexity of languages. This seemingly explains the results of a second quantitative analysis of WALS data we have also conducted, this time without considering sociopolitical factors, which suggest no trade-off (a perhaps a slight trend towards a positive correlation) between morphological and syntactic complexity across languages (Benítez-Burraco, Chen and Gil 2024).

3. A Window into phylogeny

What can the present tell us about the past? In accordance with a slightly modified version of the *Evolutionary Inference Principle for Linguistic and Cultural/Socio-Political Complexity* (Gil 2021), correlations between societal and linguistic complexity observed amongst contemporary human languages, of the sort we have highlighted above, may be used to make inferences about prior stages in linguistic phylogeny. Specifically, if particular linguistic features are found to be systematically associated with today's esoteric societies, it may be inferred that these same features were characteristic of the languages of the Upper Paleolithic era. Archaeological and paleogenetic evidence (e.g. Sikora et al., 2017, Koptekin et al., 2023) indicates that all societies were strongly esoteric at that time, with signs of exotericity increasing only recently. Invoking this principle, our findings surveyed above thus support the existence of an earlier evolutionary stage in which all languages were S-languages, with simpler syntax but more complex morphology.

4. Conclusion

That contemporary S-languages provide a model for an earlier stage in the evolution of language should not be considered surprising if we consider the effects of the social environment on language structure, and the fact that in many places, human societies still exhibit many of the sociopolitical features of Paleolithic societies. It must be kept in mind, however, that we are referring to actually observable languages, not to the linguistic abilities that underlie them. Clearly, speakers of S-languages are perfectly capable of acquiring X-languages if they are called upon to do so. In fact, in today's modern world, it is probably the case that a large majority of speakers of S-languages are also fluent in an X-language, be it a regional lingua franca or a national language.

However, some speculations in Benítez-Burraco et al. (2022) and Chen et al. (2023) point towards a deeper effect associated with the distinction between S-languages and X-languages. First, it is suggested that these two language types may make differential use of two different kinds of memory that are crucially involved in language processing: while S-languages, with their greater propensity for the kinds of irregularities typical of rich morphologies, may rely more heavily on declarative memory, X-languages, with their greater orientation towards combinatorial syntax, may tend more to call upon procedural memory. Moreover, because declarative and procedural memory seem to depend on different genes (e.g. Ullman et al., 2015), one could hypothesize this differential effect resulting in a language- type distinctive (epi)genetical signal. A more radical view would be that changes external to language resulting in the potentiation of declarative or procedural memory might have favoured the transition to the corresponding language type, S-language or X-language respectively. One such change might be the advent of more complex technologies, whose mastering would demand advanced procedural abilities. At present, this hypothesis has not yet seen any systematic empirical support, but it is a possibility we are currently testing.

Whatever the case, the results of this paper join forces with other recent studies, such as Progovac (2015), Gil (2017) and others, showing how much of the evolutionary past of human languages is still visible, in one way or another, in the contemporary linguistic landscape. Thus, linguistic typology offers a valuable window into linguistic phylogeny.

Acknowledgements

The authors would like to express their gratitude to Russell Gray and Kaius Sinnemäki for their suggestions on improving the analysis methods, as well as to the audiences of the 56th Annual Meeting of the Societas Linguistica Europaea and the 2022 JCoLE Conference for their questions and feedback. We are especially grateful to the anonymous reviewers for their constructive feedback.

References

Benítez-Burraco, Antonio, Candy Cahuana, Sihan Chen, David Gil, Ljiljana Progovac, Jana Reifegerste and Tatiana Tatarinova (2022) "Cognitive and Genetic Correlates of a Single Macro-Parameter of Crosslinguistic Variation", The Evolution of Language, Proceedings of the 14th International Conference (EVOLANG14).

Benítez-Burraco, Antonio, Sihan Chen and David Gil (2024) "The Absence of a Trade-off between Morphological and Syntactic Complexity", *Frontiers in Psychology - Language Sciences* 3:1340493.

Benítez-Burraco, Antonio and Ljiljana Progovac (2020) "A four-stage model for language evolution under the effects of human self-domestication", *Language and Communication 73.1–17.*

Bickerton, Derek (1990) *Language and Species*, University of Chicago Press.

Chen, Sihan, David Gil, Sergey Gaponov, Jana Reifegeste, Tessa Yuditha, Tatiana Tatarinova, Ljiljana Progovac and Antonio Benitez-Burraco (2023) "Linguistic and Memory Correlates of Societal Variation, A Quantitative Analysis", PsyArXiv Preprints, 10.31234/osf.io/bnz2s.

Dahl, Östen (2004) *The growth and maintenance of linguistic complexity*, John Benjamins, Amsterdam.

Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.). 2022. Ethnologue: Languages of the World. Twenty-fifth edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

Ergin, Rabia, Limor Raviv, Ann Senghas, Carol Padden and Wendy Sandler (2020) "Community Structure Affects Convergence on Uniform Word Orders: Evidence from Emerging Sign Languages", In A. Ravignani, C. Barbieri, M. Flaherty, Y. Jadoul, E. Lattenkamp, H. Little, M. Martins, K. Mudd, & T. Verhoef eds., *The Evolution of Language: Proceedings of the 13th International Conference (Evolang13),* Nijmegen, 84-86.

Gil, David (2017) "Isolating-Monocategorial-Associational Language", in H. Cohen and C. Lefebvre eds, Handbook of Categorization in Cognitive Science, Second Edition, Elsevier, Amsterdam, 471-510.

Gil, David (2021) "Tense-Aspect-Mood Marking, Language Family Size, and the Evolution of Predication", in A. Benítez-Burraco and L. Progovac eds., Prehistorical Languages, *Philosophical Transactions B.*

Gil, David and Yeshayahu Shen (2021) "Metaphors: The Evolutionary Journey from Bidirectionality to Unidirectionality", in A. Benítez-Burraco and L. Progovac eds., *Prehistorical Languages*, *Philosophical Transactions B.*

Hammarström, Harald & Forkel, Robert & Haspelmath, Martin & Bank, Sebastian. 2022. Glottolog 4.6. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://doi.org/10.5281/zenodo.6578297 (Available online at http://glottolog.org, Accessed on 2022-08-25.)

Haspelmath, Martin, Matthew Dryer, David Gil and Bernard Comrie eds. (2005) *The World Atlas of Language Structures*, Oxford University Press, Oxford.

Kirby, K. R., Gray, R. D., Greenhill, S. J., Jordan, F. M., Gomes-Ng, S., Bibiko, H. J., ... & Gavin, M. C. (2016). D-PLACE: A global database of cultural, linguistic and environmental diversity. PloS one, 11(7), e0158391.

Koplenig, Alexander (2019) "Language structure is influenced by the number of speakers but seemingly not by the proportion of non-native speakers", *Royal Society Open Science* 6(2). 181274. https://doi.org/10.1098/rsos.181274.

Koptekin, D., Yüncü, E., Rodríguez-Varela, R., Altınışık, N. E., Psonis, N., Kashuba, N., Yorulmaz, S., George, R., Kazancı, D. D., Kaptan, D., Gürün, K., Vural, K. B., Gemici, H. C., Vassou, D., Daskalaki, E., Karamurat, C., Lagerholm, V. K., Erdal, Ö. D., Kırdök, E., Marangoni, A., … Somel, M. (2023). Spatial and temporal heterogeneity in human mobility patterns in Holocene Southwest Asia and the East Mediterranean. Current biology : CB, 33(1), 41–57.e15. https://doi.org/10.1016/j.cub.2022.11.034

Lupyan, Gary and Rick Dale (2010) "Language Structure Is Partly Determined by Social Structure", *PLoS ONE* 5.1:e8559. (https://doi.org/10.1371/journal.pone.0008559.)

McWhorter, John H. (2005) *Defining Creole*, Oxford University Press, Oxford.

McWhorter, John H. (2011) *Linguistic Simplicity and Complexity, Why Do Languages Undress?* De Gruyter Mouton, Berlin.

McWhorter, John H. (2018) *The Creole Debate*, Cambridge University Press, Cambridge.

Meir, Irit, Assaf Israel, Wendy Sandler, Carol Padden and Mark Aronoff (2012) "The Influence of Community on Language Structure: Evidence from Two Young Sign Languages", *Linguistic Variation* 12.2:247-291.

Nettle, Daniel (2012) "Social scale and structural complexity in human languages", *Philosophical Transactions of the Royal Society B: Biological Sciences* 367.1597:1829-1836.

Neubauer, Simon, Jean-Jacques Hublin, and Philipp Gunz (2018) "The evolution of modern human brain shape", *Science advances*, 4(1), eaao5961 (https://doi.org/10.1126/sciadv.aao5961).

Progovac, Ljiljana (2015). *Evolutionary Syntax*. Oxford: Oxford University Press.

Progovac, Ljiljana (2019) "Gradualist Approaches to Language Evolution", in A Critical Introduction to Language Evolution, Current Controversies and Future Prospects, *SpringerBriefs in Linguistics. Cham,Switzerland: Springer. (doi:10.1007/978-3-030-03235-7_3)* 31-66.

Raviv, Limor (2020) *Language and society: How social pressures shape grammatical structure*, PhD Dissertation, Radboud University, Nijmegen.

Raviv, Limor, Antje Meyer and Shiri Lev-Ari (2019) "Larger Communities Create More Systematic Languages", *Proceedings of the Royal Society B: Vol. 286,* 907, 20191262, 17.07.2019:1-9

Raviv, Limor, Antje Meyer and Shiri Lev-Ari (2020) "The Role of Social Network Structure in the Emergence of Linguistic Structure", *Cognitive Science* 44, (DOI: 10.1111/cogs.12876).

Scerri, Eleanor M. L., and Manuel Will (2023) “The revolution that still isn't: The origins of behavioral complexity in Homo sapiens”, *Journal of human evolution*, 179, 103358 (https://doi.org/10.1016/j.jhevol.2023.103358).

Sikora, M., Seguin-Orlando, A., Sousa, V. C., Albrechtsen, A., Korneliussen, T., Ko, A., Rasmussen, S., Dupanloup, I., Nigst, P. R., Bosch, M. D., Renaud, G., Allentoft, M. E., Margaryan, A., Vasilyev, S. V., Veselovskaya, E. V., Borutskaya, S. B., Deviese, T., Comeskey, D., Higham, T., Manica, A., … Willerslev, E. (2017). Ancient genomes show social and reproductive behavior of early Upper Paleolithic foragers. Science (New York, N.Y.), 358(6363), 659–662. https://doi.org/10.1126/science.aao1807

Shcherbakova, Olena, Susanne Maria Michaelis, Hannah J. Haynie, Sam Passmore, Volker Gast, Russell D. Gray, Simon J. Greenhill, Damián E. Blasi & Hedvig Skirgård (2023) "Societies of strangers do not speak less complex languages", *Science Advances* 9(33). eadf7704. https://doi.org/10.1126/sciadv.adf7704.

Trudgill, Peter (2011) *Sociolinguistic Typology: Social Determinants of Linguistic Complexity*, Oxford University Press, Oxford.

Ullman, M. T. (2015). The declarative/procedural model: a neurobiologically motivated theory of first and second language. In Van Patten, B., and Williams, J. (Eds.), Theories in Second Language Acquisition: An Introduction (2nd ed.) (pp. 135-158). London and New York: Routledge

Wray, Allison and George W. Grace (2007) "The Consequences of Talking to Strangers: Evolutionary Corollaries of Socio-cultural influences on Linguistic Form", *Lingua* 117:543-578.

The role of gender, social bias and personality traits in shaping linguistic accommodation: An experimental approach

Lewis Ching-Yat Cheung[*1], Simon Kirby[2], and Limor Raviv[1,3]

[*]Corresponding Author: lewis.cheung@mpi.nl
[1]LEADS group, Max Planck Institute for Psycholinguistics, Nijmegen, NL
[2] Centre for Language Evolution, Department of Linguistics and English Language, University of Edinburgh, UK
[3]cSCAN, University of Glasgow, UK

In observational studies of language change, women have been shown to use more innovative forms than men, and to be more likely to adopt new variants (e.g., Chambers, 2009; Crawford, 1995; Labov, 2001). At the same time, work on vocal convergence shows a mixed pattern of results, whereby some studies find that women accommodate faster (i.e., imitating the acoustic variants of others; Namy et al., 2002), and other studies find the opposite pattern (i.e., that men accommodate faster; Pardo, 2006). However, beyond these studies, gender effects in language accommodation and diffusion of morphosyntactic variants have not been systematically tested, and it is currently not clear how language evolution may be affected by social biases and individual attributions associated with different gender groups. For example, gender effects may stem from documented differences in social attributes and personality traits between groups, such as conformity and agreeability (Weisberg et al., 2011). Here, we test how the gender and personality traits of participants, as well as the gender of their interactive partners, shape accommodation patterns in a dyadic communication experiment using an artificial language - shedding light on the role of gender in shaping language change patterns in the presence of linguistic variation.

In this pre-registered study (https://osf.io/6eudq/), following Fehér et al., (2019), we use an online communication experiment in which participants of different genders first learn how to formulate sentences using two verbs, six novel nouns (slightly altered Dutch onomatopoeia assigned to a corresponding animal picture), and a marker for plural and singular forms. During training, the plural marker is always present, while the singular marker is optional (present only 33% of the time). After learning the language, participants play a director-matcher game with a partner from either the same or different gender group (Figure 1). While participants believe they are interacting with another person, they are in fact interacting with a simulated partner. To manipulate the gender of the simulated partner, we use portraits from the FACES database (Ebner et al., 2010), and take a similar photo of each participant - creating the illusion that they are interacting with a real person at the other end. Crucially, the simulated partner always produces the singular marker which was optional during training. We examine participants' linguistic behaviour before, during, and after communication in pseudo-dyads, and test whether their tendency to accommodate to their partner (i.e., by reducing variation and increasing their use of the singular marker used by their partner) is shaped by the gender of the participant and the gender of their perceived partner. At the end of the experiment, participants complete an implicit association test (Karpinski & Hilton, 2001) and an explicit bias questionnaire (Rosencranz & McNevin, 1969; Swim et al., 1995) to assess whether they have any subconscious bias or stereotype towards different gender groups. In addition, participants fill out a self-report personality questionnaire ("Big Five"; John et al., 2008) to measure their openness to experience, extraversion, agreeableness, neuroticism, and conscientiousness.

We predict that: (1) people will be more likely to accommodate to members of their own gender group (Giles & Ogay, 2007); and (2) the likelihood of participants accommodating to their partners will be correlated with their personality traits and their attitudes/biases towards their communicative partner's gender. Critically, we predict that these patterns will be accounted for by gender-related differences in personality traits, such as women generally scoring higher than men on extraversion, agreeableness, and neuroticism. Preliminary results from N=16 participants show that while both men and women accommodated to their partner during interaction, only women persisted in using the singular marker post-interaction (with the use of the optional singular marker increasing from 25% before interaction, to 45% after interaction).

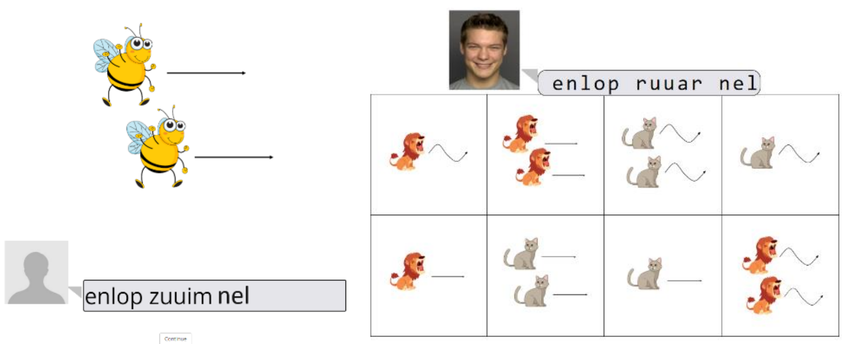


Figure 1: screenshot of the communication experiment

References

Chambers, J. K. (2009). *Sociolinguistic theory: Linguistic variation and its social significance* (Rev. ed). Blackwell.

Crawford, M. (1995). *Talking difference: On gender and language*. SAGE.

Ebner, N. C., Riediger, M., & Lindenberger, U. (2010). FACES—A database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior Research Methods*, *42*(1), 351–362. https://doi.org/10.3758/BRM.42.1.351

Fehér, O., Ritt, N., & Smith, K. (2019). Asymmetric accommodation during interaction leads to the regularisation of linguistic variants. *Journal of Memory and Language*, *109*, 104036. https://doi.org/10.1016/j.jml.2019.104036

Giles, H., & Ogay, T. (2007). Communication Accommodation Theory. In *Explaining communication: Contemporary theories and exemplars.* (pp. 293–310). Lawrence Erlbaum Associates Publishers.

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In *Handbook of personality: Theory and research*, *3rd ed.* (pp. 114–158). The Guilford Press.

Karpinski, A., & Hilton, J. L. (2001). Attitudes and the Implicit Association Test. *Journal of Personality and Social Psychology*, *81*(5), 774–788. https://doi.org/10.1037/0022-3514.81.5.774

Labov, W. (2001). *Principles of linguistic change. Vol. 2: Social factors / William Labov* (Digital print, Vol. 2). Blackwell.

Namy, L. L., Nygaard, L. C., & Sauerteig, D. (2002). Gender Differences in Vocal Accommodation: The Role of Perception. *Journal of Language and Social Psychology*, *21*(4), 422–432. https://doi.org/10.1177/026192702237958

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, *119*(4), 2382–2393. https://doi.org/10.1121/1.2178720

Rosencranz, H. A., & McNevin, T. E. (1969). A Factor Analysis of Attitudes Toward the Aged. *The Gerontologist*, *9*(1), 55–59. https://doi.org/10.1093/geront/9.1.55

Swim, J. K., Aikin, K. J., Hall, W. S., & Hunter, B. A. (1995). Sexism and racism: Old-fashioned and modern prejudices. *Journal of Personality and Social Psychology*, *68*(2), 199–214. https://doi.org/10.1037/0022-3514.68.2.199

Weisberg, Y. J., Deyoung, C. G., & Hirsh, J. B. (2011). Gender Differences in Personality across the Ten Aspects of the Big Five. *Frontiers in Psychology*, *2*, 178. https://doi.org/10.3389/fpsyg.2011.00178

Finding proportionality in computational approaches to morphological change

Alexandru Craevschi[*1,2], Sarah Babinski[1,2], and Chundra Cathcart[1,2]

[*]Corresponding Author: alexandru.craevschi@uzh.ch
[1]Department of Comparative Language Science, University of Zurich
[2]Center for the Interdisciplinary Study of Language Evolution, University of Zurich

Abstract

Proportional mechanisms are thought to play a major role in morphological change. This paper explores the extent to which simple models of morphological inflection embody proportional behavior. Using models with varying architectures and decoding schemes, we find that errors produced by these models do not often form a valid proportion with forms found in the training data. We discuss the implications of this finding for research seeking to recapitulate diachronic processes using models of this sort.

1. Introduction

Morphological paradigms change over time. Analogical change is a significant factor driving morphological shifts of this kind: when attempting to produce an inflected form, language users draw upon their knowledge of inflectional patterns from other word forms, sometimes resulting in alterations to the intended target.

Morphological paradigms can undergo restructuring through various analogical mechanisms. Traditionally, analogical change mechanisms are categorized into two types: those involving proportional and non-proportional mechanisms (Paul, 1880; Anttila, 1977; Gaeta, 2007; Hock, 2009). While paradigms may occasionally be restructured via non-proportional mechanisms (Haspelmath, 1994; Fertig, 2016; Sims-Williams, 2016), the most commonly cited types of changes affecting paradigms are analogical extension and leveling, which are typically understood to operate proportionally (Hill, 2007; Garrett, 2008). Proportional analogy refers to a phenomenon where a form or pattern is extended or generalized to create new forms or patterns within a language, maintaining consistent relationships between elements. This process involves adhering to linguistic constraints while generating both attested and unattested forms based on established patterns or paradigms. These changes fit within a framework of analogical proportions, exemplified in (1a). A proportion generates both attested and unattested forms, but for it to be considered valid, it must adhere to the linguistic constraints of the language. For instance, (1b) presents a well-formed and attested analogical proportion in Latin, where the pluralization of the second-declension Latin noun *rīvulus* is patterned after another second-declension noun *fabulus*. Conversely, (1c) demonstrates an invalid proportion where the fourth-declension noun *cēnsus* is incorrectly pluralized as **cēnsī* based on the pattern of *fabulus*; the attested plural form for *cēnsus* would be *cēnsūs* in accordance with its noun class. This disparity illustrates the importance of maintaining linguistic congruence within analogical proportions, as seen in the ill-formed proportion (1d) which attempts to apply a feminine form ending in *-a* to generate the plural of a masculine noun ending in *-us*.

(1) a. A : B :: C : *x*
 b. fabulus : fabulī :: rīvulus : **rīvulī**
 c. fabulus : fabulī :: cēnsus : ***cēnsī**
 d. fābula : fābulae :: rīvulus : ***rīvulae**

The extent to which computational models of morphological change exhibit proportional behavior remains unexplored. Earlier computational work on morphological learning exploits pairwise relationships between inflected forms in order to establish proportional bases for generating inflectional forms (Neuvel & Fulop, 2002). However, the role of proportionality in neural models of inflection, which learn linear and/or nonlinear mappings between semantic features and phonological cues, is not fully understood. Linear discriminative learning (LDL) (Baayen, Chuang, & Blevins, 2018; Baayen, Chuang, Shafaei-Bajestan, & Blevins, 2019) is a framework which maps meaning to form and vice versa by learning linear relationships between vector semantic and phonological cues. Its proponents argue that LDL generalizes the standard four-part analogy (1a) beyond set-based conceptions of semantics (e.g., {DOG, SINGULAR}) to vector semantic representations representing the collocational distributions in which a form is found.

In this study, we explore the extent to which proportional behavior emerges in computational models of morphological inflection without the models being explicitly coded to use four-part analogies from (1a) in the process of inference. We apply models of morphological inflection to morphological data sets from different languages, allowing different properties of the models used, namely the architecture and decoding schemes, to vary across model settings. We employ an algorithm to find proportions in the training data that support attested and predicted forms in the test data. Models with a linear regression-based architecture perform consistently better than Long Short-Term Memory (LSTM) models in terms of rates of proportional errors. Proportional support for a test form in the training data is a significant predictor of whether or not a morphological inflection model will generate it accurately, though this can be interpreted as a proxy for type frequency. Analyses of the errors produced by the models show an overwhelmingly low degree of proportionality. Our results suggest that if changes in morphological paradigms are overwhelmingly proportional, then computational models of morphological learning should be used with care when simulating historical changes.

2. Data

Verbal paradigms were sourced from Unimorph (McCarthy et al., 2020) and converted to a broad IPA transcription using Epitran (Mortensen, Dalmia, & Littell, 2018) for most languages, with a few manual corrections. Phonemic transcriptions for English were taken from the Carnegie Mellon Pronouncing Dictionary (Rudnicky, 2015), available through the Natural Language Toolkit (Bird, Klein, & Loper, 2009). The glosses available in UniMorph, as well as lemmas, were converted to one-hot representations of inflectional features. Models were applied to data from the following languages: Arabic, Dutch, English, Italian, German, Polish, Portuguese, Russian, and Spanish. The main criteria for selection were (1) availability of verbal paradigms in UniMorph; and (2) availability of a grapheme-to-phoneme (g2p) system for obtaining phonological representation of forms. To alleviate the problem of different numbers of lemmas available per language and also to avoid extreme processing times for some of the data sets, we limited ourselves to a sample of 500 lemmas per language or just used all the lemmas, in case a language has fewer than 500 verbal lemmas (e.g., Zulu). Data and code are available at `https://gitlab.uzh.ch/chundra.cathcart/evolang_2024`.

3. Methods

We evaluate the performance of four models varying in the way meaning is mapped onto form and the way predicted sequences are generated. We probe the extent to which errors produced are supported by a proportional basis in the training data, and explore other properties of proportionality with respect to model performance.

Our models vary in the way they map meaning onto form. Following Baayen et al. (2018, 2019), in one set of models we use linear regression to learn linear mappings between inflectional features and trigram phoneme sequences, which can be used to predict phoneme sequences from inflectional features. Linear regression models were fitted using ordinary least squares.

Another set of models utilizes Long Short-Term Memory (LSTM) neural networks to introduce non-linearity. These models follow a standard encoder-decoder architecture commonly employed in sequence-to-sequence tasks. The architecture consists of two embedding layers, one for inflectional features and another for phonemic form. The inflectional features' embedding is fed into the LSTM encoder, with both embedding and hidden layer dimensions set to 128. The output of the encoder is then passed to the decoder for generation of the phonological form. LSTM models were implemented in Keras (Chollet et al., 2015) and trained with the Adam optimizer (Kingma & Ba, 2015) with a categorical cross-entropy loss function. The models were stopped early once overfitting on validation data was observed.

Finally, we vary the models with two different methods for sequence generation of predicted forms at the inference stage: beam search and greedy decoding. Greedy decoding selects the most probable token at each step, in this case selecting the most probable initial trigram/phoneme given some inflectional features before moving on to the following trigram/phoneme. Beam search, on the other hand, maintains a set of top-N candidates, exploring multiple possibilities simultaneously. We considered the top 2 most probable candidates at each generation step, ultimately picking out the sequence with the highest probability overall. For linear regression models decoded using beam search, we follow Baayen et al. (2018) in training a second model that maps trigram phoneme sequences to semantic vectors, choosing the candidate sequence whose predicted semantics correlates most strongly with the input semantic vector.

Models are run separately for each language. To evaluate model performance, we carry out K-fold cross-validation ($K = 10$), randomly holding out 10% of the forms in each data set used as test data. We vary the random number seeds used to sample lemmas and generate folds, using 5 different seeds for each stochastic dimension.

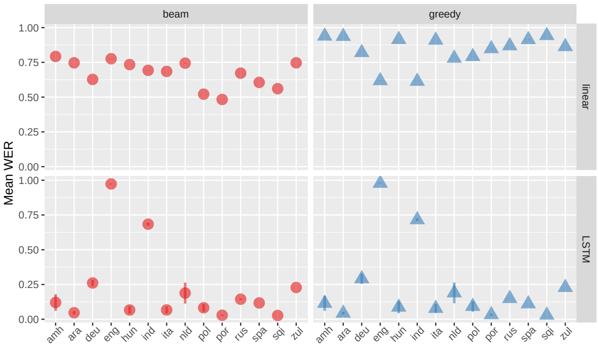


Figure 1. Mean word error rate (WER) by language for each model setting. Error bars (where visible) represent variance across different folds and random number seeds.

4. Model Results

We assess model performance according to the word error rate (WER, the proportion of test items that are produced with at least one error) and the phoneme error rate (PER, the mean normalized Levenshtein distance between each target and each predicted form). WER and PER values are displayed in Figures 1–2.

WER values are relatively high for linear models. Results from LSTM models show considerably lower WER values, with the exception of English and Indone-

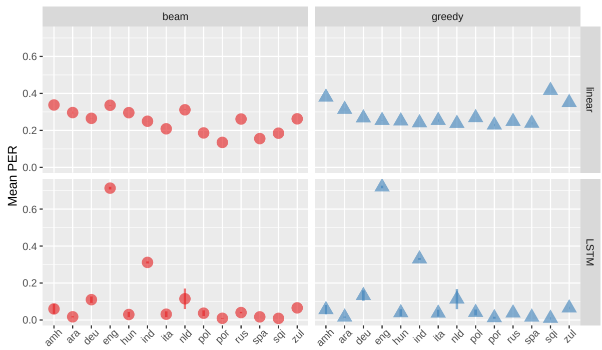


Figure 2. Mean phoneme error rate (PER) by language for each model setting. Error bars (where visible) represent variance across different folds and random number seeds.

sian. The poor performance for these languages is striking, and may be due to their generally smaller paradigm size in comparison to the other languages, which are morphologically richer. PER values display a similar trend, although the difference between LSTM and linear regression models is significantly lower in case of PER.

Beam search and greedy decoding schemes do not show consistent differences from each other in terms of performance for these two error metrics. Certain languages show better results in greedy decoding with others benefiting more from beam search. In LSTM models, the differences between beam search and greedy decoding are almost nonexistent.

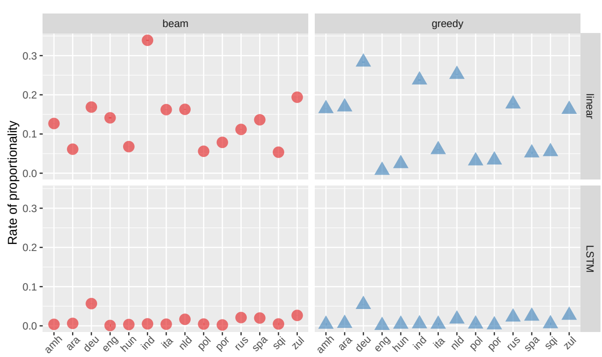


Figure 3. Proportion of model errors with support from at least one proportional basis, by language and for each model setting.

5. Proportionality

We identify proportions in the training data that could give rise to forms in the test data (following Lepage, 1998; Sims-Williams, 2016, 2021). For a given paradigm cell c_i in a given lemma ℓ found in a test split, we iterate through all lemmas $l \in \boldsymbol{\ell}_{\text{training}}$ in the training data, and for each cell $c_j \neq c_i$, generate the proportion (l, c_j):(l, c_i)::(ℓ, c_j):x. We tabulate the number of proportions in the training data that support each attested target form as well as each predicted form. Using a mixed-effects logistic regression model with word error rate (with values of 1 representing errors) as a response variable and log-transformed proportional strength as a predictor with random intercepts and slopes by language, architecture, and decoding scheme, we find that the proportional support for an attested target form is a significant predictor of accuracy, though the effect is weak ($\beta = -0.0055$ $p <$ .001). This may not indicate anything interesting about the effect of proportionality on model accuracy, but may have to do more generally with type frequency. We compute the proportion of errors for each model for which at least one proportion is available in the training data. These values are displayed in Figure 3. Linear architectures generate more proportional errors than LSTM. In many cases, these errors involve regularization, in which case the incorrect prediction will have more proportional support than the attested target form. Greedy decoding and beam search appear to have little influence on the rates of proportionality.

6. Discussion

This paper explored the performance of different models of morphological inflection, with an eye to assessing the extent to which models exhibit proportional behavior. We find that errors produced by these models are unlikely to have support from proportional bases in the training data, with under 35% of errors found across all model settings. Models making use of linear mappings between semantic and phonological cues are found to generate a higher degree of proportional errors.

Our results have implications for research that aims to simulate historical changes using computational models of morphological inflection (e.g., Cotterell, Kirov, Hulden, & Eisner, 2018). If analogical changes that restructure morphological paradigms are in fact overwhelmingly proportional, then care is warranted when choosing models for this particular task. Even for models from the framework of linear discriminative learning, which in a sense incorporates proportionality by learning linear mappings between phonological sequences and semantic variables, the degree of proportional errors produced depends on a range of factors and displays variability across languages. Future work will benefit also from exploring the degree to which models of this sort generate proportions traditionally thought to be invalid, such as *four*:*fork*::*three*:*threek*, or *ear*::*hear*:*eye*::*heye* (Kiparsky, 1968; Deutscher, 2002), in order to probe the extent to which such models can be used to reliably recapitulate processes of diachronic change.

References

Anttila, R. (1977). *Analogy.* The Hague: Mouton.

Baayen, R. H., Chuang, Y.-Y., & Blevins, J. P. (2018). Inflectional morphology with linear mappings. *The Mental Lexicon*, *13*(2), 230–268.

Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., & Blevins, J. P. (2019). The Discriminative Lexicon: A Unified Computational Model for the Lexicon and Lexical Processing in Comprehension and Production Grounded Not in (De)Composition but in Linear Discriminative Learning. *Complexity*, *2019*.

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit.* O'Reilly Media, Inc.

Chollet, F., et al.. (2015). *Keras.* `https://keras.io`.

Cotterell, R., Kirov, C., Hulden, M., & Eisner, J. (2018). On the diachronic stability of irregularity in inflectional morphology. *arXiv preprint arXiv:1804.08262*.

Deutscher, G. (2002). On the misuse of the notion of 'abduction' in linguistics. *Journal of Linguistics*, *38*(3), 469–485.

Fertig, D. (2016). Mechanisms of paradigm leveling and the role of universal preferences in morphophonological change. *Diachronica*, *33*(4), 423–460.

Gaeta, L. (2007). Is analogy economic? In G. B. Fabio Montermini & N. Hathout (Eds.), *Selected Proceedings of the 5th Décembrettes: Morphology in Toulouse.* Somerville, MA: Cascadilla Press.

Garrett, A. (2008). Paradigmatic uniformity and markedness. In J. Good (Ed.), *Explaining linguistic universals: Historical convergence and universal grammar* (pp. 124–143). Oxford: Oxford University Press.

Haspelmath, M. (1994). The growth of affixes in morphological reanalysis. In *Yearbook of Morphology 1994* (pp. 1–29). Springer.

Hill, E. (2007). Proportionale Analogie, paradigmatischer Ausgleich und Formerweiterung: ein Beitrag zur Typologie des morphologischen Wandels. *Diachronica*, *24*(1), 81–118. (Publisher: John Benjamins)

Hock, H. H. (2009). *Principles of historical linguistics.* Walter de Gruyter.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *The 3rd International Conference for Learning Representations.* San Diego.

Kiparsky, P. (1968). Linguistic universals and linguistic change. In E. Bach & R. Harms (Eds.), *Universals in linguistic theory* (pp. 170–202). Holt, Rinehart, and Winston.

Lepage, Y. (1998). Solving analogies on words: an algorithm. In *The 17th International Conference on Computational Linguistics (COLING)* (Vol. 1).

McCarthy, A. D., Kirov, C., Grella, M., Nidhi, A., Xia, P., Gorman, K., Vylomova, E., Mielke, S. J., Nicolai, G., Silfverberg, M., Arkhangelskiy, T., Krizhanovsky, N., Krizhanovsky, A., Klyachko, E., Sorokin, A., Mansfield, J., Ernštreits, V., Pinter, Y., Jacobs, C. L., Cotterell, R., Hulden, M., & Yarowsky, D. (2020). UniMorph 3.0: Universal Morphology. In *Proceed-*

ings of the Twelfth Language Resources and Evaluation Conference (pp. 3922–3931). Marseille, France: European Language Resources Association.

Mortensen, D. R., Dalmia, S., & Littell, P. (2018). Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).*

Neuvel, S., & Fulop, S. A. (2002). Unsupervised learning of morphology without morphemes. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning* (Vol. 6, pp. 31–40).

Paul, H. (1880). *Prinzipien der Sprachgeschichte.* Halle: Niemeyer.

Rudnicky, A. (2015). *The Carnegie Mellon Pronouncing Dictionary, Version 0.7b.*

Sims-Williams, H. (2016). *Analogy in morphological change.* Unpublished doctoral dissertation, University of Oxford.

Sims-Williams, H. (2021). Token frequency as a determinant of morphological change. *Journal of Linguistics*, 1–37.

Testing the Linguistic Niche Hypothesis in Large Language Models with a Multilingual Wug Test

Anh Dang[1,2,*], Limor Raviv[1,3], and Lukas Galke[1]

[*]Corresponding Author: thithaoanh.dangthithaoanh@ru.nl
[1]LEADS group, Max Planck Institute for Psycholinguistics, Nijmegen, NL
[2]CLS, Radboud University, NL
[3]cSCAN, University of Glasgow, UK

The linguistic niche hypothesis states that languages spoken by larger societies tend to have less complex morphological systems (Lupyan & Dale, 2010), which may be caused by a learnability advantage of L2 learners for less complex systems (Wray & Grace, 2007; Hudson Kam & Newport, 2009). Despite the high impact of this theory on the field of language evolution and adaptation (Gibson et al., 2019; Bentz et al., 2015, 2018; Lupyan & Dale, 2016), recent studies (Koplenig et al., 2023; Shcherbakova et al., 2023) challenge the linguistic niche hypothesis and suggest an opposite relationship between morphological complexity and population size, whereby larger societies actually have more complex morphological systems. Here, we test the underlying assumption that languages with less complex morphological systems are easier to learn for language models. To this end, we evaluate to what extent morphological generalization is influenced by linguistic complexity and population size in a new type of learner: large language models (LLMs). Testing cross-linguistic patterns of language learning in LLMs trained on large amounts of human-generated text is particularly interesting given recent findings highlighting the similarity between humans and such models with respect to language learning and processing (Galke et al., 2023; Webb et al., 2023; Srikant et al., 2022) and to the emergence of syntactic structure within the model's learned attention patterns (Manning et al., 2020). While there is little cross-linguistic work on the morphological knowledge of LLMs in relation to the degree of morphological structure, some work suggests that LLMs often fail to generate the correct inflected forms of words that are not present in the training data, regardless of the size of the training set and the target language (Liu & Hulden, 2022). As such, it is currently unknown to what extent LLMs can learn to generalize their morphological knowledge and to what extent their generalization capabilities are affected by the degree of linguistic complexity in their input.

In our study, we developed a multilingual version of the Wug Test, an artificial word completion test that is typically used to evaluate the inflectional and derivational morphological knowledge of children (Berko, 1958), and applied it to the

GPT family of large language models (Brown et al., 2020; Ouyang et al., 2022). We considered six different languages, namely German, Vietnamese, Spanish, French, Romanian, and Portuguese, which vary in their degree of morphological complexity and well as the amount of text available for them. For each language, we first asked GPT-4 to translate the questions from the original Wug Test – translations that were then evaluated and corrected by native speakers. Then, LLMs were provided with the translated questions (i.e., a sentence in which the fantasy word, e.g., 'wug', represents either a noun or a verb), and were made to respond with the inflected form (e.g,. plural form, past tense). Since the fantasy words (very likely) do not exist in the respective training data, the models needed to use their morphological knowledge of the language in order to be successful. The models' answers were then evaluated by native speakers, who judged whether the generated inflected and derived forms conform to their native language's morphological rules (see Additional File for examples).

To connect our results with the linguistic niche hypothesis, we test whether accuracy was predicted by morphological complexity and training size, taking into account Ackerman and Malouf (2013) distinction between e-complexity (the number of rules and irregularities) and i-complexity (how well are morphemes predicted by their context). E-complexity was measured using Lupyan and Dale (2010)'s original complexity scores (LNH in the table), and i-complexity was measured using Bentz et al. (2015)'s lexical diversity score, based on Shannon entropy (H_{scaled}).

Language	**%train**	**LNH**	$\mathbf{H}_{\text{scaled}}$	**Model**	**Correct**	**Unclear**	**Wrong**
German	1.68%	-12	0.4648	GPT-3.5	66%	5%	29%
				GPT-4	62%	5%	33%
Vietnamese	0.03%	-16	-1.2099	GPT-3.5	71%	0%	19%
				GPT-4	81%	0%	19%

The table shows the results for German and Vietnamese. We find that while both GPT-3.5 and GPT-4 were generally capable of generating the correct inflected forms for unknown words, they were not always able to inflect them correctly. Notably, our initial results are promising: Despite German having 50 times more representation than Vietnamese in GPT-3's training data (1.67583% compared to 0.03373%), the model scores higher on the less complex (w.r.t. LNH and H_{scaled}) Vietnamese morphological system – indicating that less complex morphological systems are learned better by LLMs, even given much less data. Our findings thus provide a first indication that multilingual LLMs satisfy the underlying assumption of the linguistic niche hypothesis – i.e., that languages with more complex morphologies are harder to learn.

References

Ackerman, F., & Malouf, R. (2013). Morphological organization: The low conditional entropy conjecture. *Language*, *89*(3), 429–464.

Bentz, C., Dediu, D., Verkerk, A., & Jäger, G. (2018). The evolution of language families is shaped by the environment beyond neutral drift. *Nature Human Behaviour*, *2*(11), 816–821.

Bentz, C., Verkerk, A., Kiela, D., Hill, F., & Buttery, P. (2015). Adaptive Communication: Languages with More Non-Native Speakers Tend to Have Fewer Word Forms. *PLOS ONE*, *10*(6), e0128254.

Berko, J. (1958). The child's learning of english morphology. *Word*, *14*(2-3), 150–177.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 1877–1901). Curran Associates, Inc.

Galke, L., Ram, Y., & Raviv, L. (2023, February). *What makes a language easy to deep-learn?* (No. arXiv:2302.12239). arXiv.

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How Efficiency Shapes Human Language. *Trends in Cognitive Sciences*, *23*(5), 389–407.

Hudson Kam, C. L., & Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology*, *59*(1), 30–66.

Koplenig, A., Wolfer, S., & Meyer, P. (2023). A large quantitative analysis of written language challenges the idea that all languages are equally complex. *Scientific Reports*, *13*(1), 15351.

Liu, L., & Hulden, M. (2022). Can a Transformer Pass the Wug Test? Tuning Copying Bias in Neural Morphological Inflection Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 739–749). Dublin, Ireland: Association for Computational Linguistics.

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PLoS ONE*, *5*(1), e8559.

Lupyan, G., & Dale, R. (2016). Why Are There Different Languages? The Role of Adaptation in Linguistic Diversity. *Trends in Cognitive Sciences*, *20*(9), 649–660.

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, *117*(48), 30046–30054.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Shcherbakova, O., Michaelis, S. M., Haynie, H. J., Passmore, S., Gast, V., Gray, R. D., Greenhill, S. J., Blasi, D. E., & Skirgård, H. (2023). Societies of strangers do not speak less complex languages. *Science Advances*.

Srikant, S., Lipkin, B., Ivanova, A. A., Fedorenko, E., & O'Reilly, U.-M. (2022). Convergent Representations of Computer Programs in Human and Artificial Neural Networks. In *Advances in Neural Information Processing Systems.*

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 1–16.

Wray, A., & Grace, G. W. (2007). The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua*, *117*(3), 543–578.

Starlings recognize simple dependency patterns

Emily Davis [*1], Robert Kluender[1], and Timothy Q. Gentner[2]

[*]Corresponding Author: e4davis@ucsd.edu
[1]Linguistics, UC San Diego, California, US
[2]Psychology, UC San Diego, California, US

Long-distance dependencies between elements are an important aspect of human language. There is some experimental evidence that animals, including songbirds, are capable of learning to recognize and generalize dependency patterns in nonlinguistic auditory or visual sequences. Here we present the results of an experiment following up on the important work of Gentner et al. (2006) and Abe & Watanabe (2011), who determined that songbirds were capable of learning to recognize some types of center-embedded patterns in auditory sequences. We tested the ability of three starlings to learn to recognize and generalize another type of simple center-embedded pattern, of the form A-B-A, and we report on one bird's successful performance. These results, like those of previous experiments, suggest that the ability of animals to learn particular dependency patterns shows individual variation.

1. Introduction: Sequence learning by animals and its significance

One of the major debates in language evolution has centered around the question of whether particular aspects of the language faculty, such as long-distance dependencies, are unique to human language and cognition. This question of uniqueness may be addressed by testing the ability of animals to recognize analogous patterns in nonlinguistic sequences (Fitch & Friederici 2012).

Long-distance dependencies are of particular interest because they are a prerequisite to recursive center-embedding, i.e. the nesting of one dependency within another of the same type (Rogers & Pullum 2006, Rohrmeier et al. 2015), and therefore are related to the vital and controversial issue of recursion (Hauser et al. 2002). Dependencies in language, e.g. between a subject and a verb, are determined by meaning, but in nonlinguistic sequences, a dependency may consist of two identical or similar elements (Gebhart et al. 2009, Dedhe et al. 2023), or a pair of elements that consistently co-occur (de Vries 2008). Two simple forms of nonlinguistic sequences with long-distance dependencies are the form A^nB^n (e.g. AABB, AAABBB), in which the A and B elements can be analyzed as forming embedded "bracket" pairs, and the form *ABA*, in which the A's are matched elements and the B a string of one or more different intervening elements. Some animals, particularly passerine birds, have demonstrated the ability to recognize and generalize auditory sequences of these types, and even more complex sequences with multiple center-embedded dependencies. This paper reports on a new experiment concerning the ability of European starlings to recognize and generalize an auditory pattern of the form ABA. Starlings are of particular interest because of their complex songs and vocal learning abilities, and are excellent problem solvers (Audet et al. 2023).

Previous experiments have demonstrated the ability of starlings and other songbirds, as well as nonhuman primates (Jiang et al. 2018, Ferrigno et al. 2020), to recognize dependency patterns, though the interpretation of the results is not always clear. Gentner et al. (2006) found that some European starlings could learn to distinguish sequences of the form A^nB^n from strings not matching the pattern (e.g. ABAB, AABBB). However, it is uncertain whether such sequences are most accurately analyzed as exhibiting center-embedded dependencies between A and B elements (Rogers and Pullum 2011: 339); they may also be recognized by a count-and-match process. Therefore, it remains uncertain whether these results actually demonstrate an ability to track dependencies between elements. Similarly, Van Heijningen et al. (2009) showed that zebra finches could learn to recognize A^nB^n patterns in sequences of motifs from their songs. However, most of their birds did not succeed at this task, and further probe tests suggested that the one bird that succeeded was using simpler processes to solve the task, e.g. the presence of adjacent identical motifs.

In another follow-up to Gentner et al. (2006), Abe and Watanabe (2011) demonstrated that Bengalese finches were able to learn both simple dependency patterns and more complex center-embedded patterns. In their first experiment, the finches were familiarized with auditory sequences that contained a single long-distance dependency. In a second experiment, the birds were successfully trained to recognize patterns with multiple center-embedded dependencies, e.g. ABCBA, ABCDCBA. The birds learned to recognize these sequence types consistently, distinguish them from other sequence types, and generalize to new sequences following the same patterns. However, Beckers et al. (2012) argued that the finches could have accomplished the task in the second experiment by memorizing substrings, rather than generalizing the abstract center-embedded pattern. Still, the results of Abe & Watanabe's (2011) prior experiment, in which birds recognized long-distance dependencies, are robust and intriguing, since these single dependencies are a necessary precondition for deeper center-embedding patterns. Here, we report on an auditory task with starlings which tested their ability to recognize A-B-A patterns.

2. The current experiment

The present experiment attempted to test whether European starlings could learn to distinguish simple dependency patterns of the A-B-A form from patterns with the same elements in a different order (A-A-B). These patterns unambiguously display center-embedding, i.e. a B element between two matching A's, but they may also be amenable to other forms of pattern recognition, e.g. the presence or absence of adjacent identical elements. Once a bird had learned to classify multiple sequences following this pattern, we used probe stimuli to determine whether it was truly generalizing the A-B-A pattern or using other cues.

2.1. Methods

2.1.1. Subjects and stimuli

Data were collected from three wild-caught adult starlings. All stimuli were artificially concatenated sequences of 3 or 4 motifs from recorded starling songs (as in Gentner et al. 2006), each 800-1000ms long, separated by 200ms silences.

In the first phase of the task, the bird was trained to distinguish two specific sequences of motifs: A-B-A and A-A-B, i.e. a sequence with a non-adjacent dependency between the two identical elements, and one in which those elements were adjacent. Subsequent phases added new training strings; the types of stimuli used at each phase, and the subjects' performance, are detailed below.

2.1.2. Experimental setup and trials

Each bird was housed individually in an acoustically isolated operant conditioning chamber, with a feeding apparatus activated by performing trials. The bird was able to initiate trials at any time by inserting its beak into the central peck-port. After doing this, the bird heard a stimulus and was required to respond within a two-second window. The stimuli were coded as either go-right or go-left, and pecking the correct port for a given stimulus type resulted in a food reward. Test trials added in the testing phases (see Section 2.2) were initiated the same way as normal trials, i.e. by the bird pecking the central port. These trials employed random reinforcement: whether the bird's response was correct or incorrect, a reward would be given 50% of the time. A session (used to measure the birds' performance over time) was defined as a block of 100 normal trials, plus 100 test trials in the testing phases; all stimuli were presented in random order. Performance in a session was measured as percentage of correct responses on all normal trials, disregarding trials in which the bird did not respond within two seconds. Performance on test trials was analyzed separately and compared to normal trial performance, as discussed below.

2.2. Results by phase

2.2.1 Training 1-2: ABA and AAB strings

In the first training phase (Training 1), the bird was trained to distinguish two strings using the trial structure detailed above: ABA (go left) and AAB (go right). Once the bird was performing stably above chance (10 consecutive sessions above 65%) on these two strings, the additional sequences BAB and BBA were introduced to the set of stimuli (Training 2). Once performance was stable above chance (20 sessions above 65%) on all four strings, the first testing phase began. Table 1 lists the training and testing phases, and the strings introduced in each phase. The strings introduced in subsequent phases, and the bird's performance, are described in following subsections.

Two of the three subjects, Birds 1 and 3, were not successful on the first phase of the task; i.e. performance on the first four trained strings was not significantly greater than chance after 300 sessions. However, Bird 2 displayed above-chance performance at 200 sessions and was therefore advanced to the testing phases. From here on we will track the performance of Bird 2.

Table 1: Training and test stimuli for each phase of task

Phase	*Left*	*Right*
Training 1	ABA	AAB
Training 2	ABA BAB	AAB BBA
Testing 1	ABA BAB + *CDC DCD*	AAB BBA + *CCD DDC*
Training 3	ABA BAB CDC DCD	AAB BBA CCD DDC
Testing 2	ABA BAB CDC DCD + *EFE FEF*	AAB BBA CCD DDC + *EEF FFE*
Training 4	ABA BAB CDC DCD EFE FEF	AAB BBA CCD DDC EEF FFE
Testing 3	ABA BAB CDC DCD EFE FEF + *ABBA CBBC*	AAB BBA CCD DDC EEF FFE + *ABBC CBBA*
Testing 4	ABA BAB CDC DCD EFE FEF + *AAA BBB CCC DDD EEE FFF*	AAB BBA CCD DDC EEF FFE
Testing 5	ABA BAB CDC DCD EFE FEF	AAB BBA CCD DDC EEF FFE + *ABC DEF*

2.2.2 Testing 1-2 and Training 3-4: Generalization of the ABA/AAB pattern

In testing phases 1 and 2, two successive sets of test strings were introduced which followed the same ABA/AAB pattern but with new motifs (see Table 1). In each test phase, the new test strings were randomly reinforced (as defined in Section 2.1.2), in order to evaluate whether the bird was generalizing the pattern it had learned to new strings. In the following training phases (3 and 4), the test strings for the prior test phase were reinforced. When performance on all training strings was stably above chance (at least 10 sessions above 65%), the next test phase commenced, and so on. In contrast to the 200 sessions it took to learn the original pattern, Bird 2 immediately generalized to new strings within the ABA/AAB paradigm, i.e. in each test stage, performance on the introduced test strings was significantly above chance (i.e. 65% or higher for 10 sessions) and did not differ significantly from performance on previously learned strings (Testing 1: $t(52) = 1.36$, $p = 0.178$; Testing 2: $t(9) = 0.861$, $p = 0.41$). Once the bird's performance was stable on all these strings, we moved on to testing two hypotheses for which criteria it was using. While persistent correct performance might be taken to indicate learning and generalization of a simple ABA embedded pattern, this is not the only possibility. Correct performance on the 12 learned strings could be accounted for by an alternative rule: recognition of strings containing an adjacent pair of identical motifs (e.g. BB). We tested these two hypotheses with three sets of test strings that fulfilled both criteria, or neither, in order to infer what criterion the bird was using to classify strings.

2.2.3. Testing 3: Disambiguating the learned pattern

This stage was intended to disambiguate which of two strategies the bird might have successfully used to classify the twelve trained strings, using four test strings that all contained two adjacent identical motifs: ABBA, CBBC, ABBC, and CBBA. Consistently classifying ABBA and CBBC as go-left, and the others as go-right, would mean that the bird was generalizing the first/last match rule. Conversely, going right on all strings containing a doubled motif would produce a pattern of largely “correct” performance on non-match strings and poor performance on the others; e.g. ABBA would be classified incorrectly as go-right while ABBC would be correctly placed in the same category. Performance on these stimuli, however, was around chance and did not consistently show either pattern. The bird’s performance on these four strings was significantly lower than performance on the twelve learned strings at the time ($t(49) = 2.37$, $p = 0.02$). Neither of the hypothesized patterns (go left for first-last match, or go right for two adjacent identical motifs) clearly emerged for individual strings. The subsequent phases attempted to clarify further which strategy the bird was using.

2.2.4. Testing 4: Strings that fit both criteria

The intention of Testing 4 was to further clarify whether the bird was classifying strings by one of the two alternative strategies discussed previously. This was performed with strings that fit both criteria: three adjacent identical motifs, e.g. AAA, BBB. If the bird was using the first-last match criterion, AAA-type strings would be classified as go-left; if its criterion was two adjacent identical motifs, they would be classified as go-right. No significant difference was observed between performance on AAA-type strings and the learned ABA/AAB stimuli ($t(18) = 0.17$, $p = .86$). Over 10 sessions with these test strings, the bird tended toward going left at above chance rates, suggesting it was in fact generalizing the intended pattern.

2.2.4. Testing 5: Strings that fit neither criterion

While the bird’s classification of AAA-type strings suggested it was successfully generalizing the first-last-match pattern, we probed it further using strings that exemplified neither pattern: strings like ABC with no repeat elements. If the bird was classifying strings based on first-last match, it would be expected to categorize these strings as not representing this pattern (i.e. go right); on the other hand, if it was classifying strings based on adjacent identical elements, it would place them in the same category as the first-last match strings (i.e. left). In this case, the bird preferred to go left on these strings ($t(19) = 6.7$, $p < 0.001$). This suggests it may have been classifying them based on the absence of two adjacent identical elements, a feature for which it had been previously trained to go right. Alternatively, it may have been uncertain how to respond to these strings, as it had not previously heard any stimuli with no repeated elements.

2.3. Discussion and conclusion

Our results with the successful bird contribute to the evidence that avians can learn to recognize dependency patterns in auditory sequences. Once this bird learned the pattern exemplified by the initial four sequences (ABA / BAB vs. AAB / BBA), it immediately generalized to new sequences exemplifying the same patterns with different elements, demonstrating that it had not merely memorized the sequences it was first trained on but learned the relevant pattern.

However, the results of the later test phases are not entirely clear as to whether the successful bird was relying on a strategy involving dependencies rather than a more “local” strategy, i.e. listening for adjacent identical motifs. In the case of the strings ABBA and CBBC, and their non-match counterparts, the bird’s performance did not show a very clear tendency toward classifying strings based either on two adjacent identical motifs, or on a match between first and last motif. The results from AAA-type strings were clear: the bird classified these as go-left, implying it had generalized the first-last pattern rather than the two adjacent motifs pattern, although they could also have been classified as go-right based on the presence of two adjacent identical motifs. Strings with three different motifs, e.g. ABC, were predominantly classified as go-left, suggesting a tendency to go right on strings with two adjacent identical motifs, and left otherwise. The bird may have been attentive to both patterns as criteria for classifying strings because it had been very extensively trained to distinguish ABA and AAB patterns, and may have had difficulty with the ABC-type strings because it had not previously been exposed to stimuli with no repeats. In general, however, the bird’s responses on test strings suggest a correct generalization of the first-last-match pattern.

As in previous animal sequence-learning studies (e.g. Gentner et al. 2006, Van Heijningen et al. 2009, Jiang et al. 2018, Ferrigno et al. 2020, Liao et al. 2022), not all of the animal participants were equally successful at the basic task. In this case only one bird learned the pattern. Two birds out of the three initially tested did not achieve consistently higher than chance performance on the first four strings (ABA/BAB and AAB/BBA) after around 300 sessions. The bird that did succeed learned to generalize this pattern after extensive training. These results suggest that among starlings, as among other nonhuman species, ability on center-embedding pattern recognition tasks can vary greatly across individuals. This may reflect different aptitudes for particular patterns, or other cognitive factors.

References

Abe, K., & Watanabe, D. (2011). Songbirds possess the spontaneous ability to discriminate syntactic rules. *Nature Neuroscience, 14*(8), 1067.

Audet, J. N., Couture, M., & Jarvis, E. D. (2023). Songbird species that display more-complex vocal learning are better problem-solvers and have larger brains. Science, 381(6663), 1170-1175.

Chomsky, N. & Miller, G. (1963). Introduction to the formal analysis of natural languages. In Luce, R., Bush, R., & Galanter, E. (eds.) *Handbook of Mathematical Psychology, Vol 2.* New York: Wiley, 269-323.

Dedhe, A. M., Piantadosi, S. T., & Cantlon, J. F. (2023). Cognitive Mechanisms Underlying Recursive Pattern Processing in Human Adults. *Cognitive Science, 47*(4), e13273.

De Vries, M., Monaghan, P., Knecht, S., & Zwitserlood, P. (2008). Syntactic structure and artificial grammar learning: The learnability of embedded hierarchical structures. *Cognition, 106*, 763-774

Ferrigno, S., Cheyette, S.J., Piantadosi, S.T., and Cantlon, J.F. (2020). Recursive sequence generation in monkeys, children, U.S. adults, and native Amazonians. *Science Advances, 6*(26).

Gebhart, A.L., Newport, E.L., and Aslin, R.N. (2009). Statistical learning of adjacent and nonadjacent dependencies among nonlinguistic sounds. *Psychon. Bull. Rev., 16*(3):486-490. doi:10.3758/PBR.16.3.486

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: what is it, who has it, and how did it evolve?. *Science*, *298*(5598), 1569-1579.

Jiang, X., Long, T., Cao, W., Li, J., Dehaene, S., & Wang, L. (2018). Production of supra-regular spatial sequences by macaque monkeys. *Current Biology, 28*(12), 1851-1859.

Liao, D. A., Brecht, K. F., Johnston, M., & Nieder, A. (2022). Recursive sequence generation in crows. *Science Advances*, *8*(44).

Pullum, G. K., & Rogers, J. (2006). Animal pattern-learning experiments: Some mathematical background. Ms. Radcliffe Institute for Advanced Study/Harvard University.

Rogers, J., & Pullum, G. K. (2011). Aural pattern recognition experiments and the subregular hierarchy. *Journal of Logic, Language and Information, 20*, 329-342.

Van Heijningen, C. A., De Visser, J., Zuidema, W., & 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, 10*6(48), 20538-20543.

Fixation Times for Language Evolution in Social Networks

Bart de Boer[*1]

[*]bart@ai.vub.ac.be
[1]AI-lab, Vrije Universiteit Brussel, Brussels

1. Introduction

Biological and cultural evolution both play a role in understanding evolution of language, but their interaction is complicated by their great difference in speed (Kirby & Hurford, 1997). This has led to debate on what could have evolved biologically (Baronchelli et al., 2013; Christiansen & Chater, 2008; De Boer & Thompson, 2018) with many of the arguments based on computational and mathematical analyses. However, no formal model to predict the speed of cultural evolution appears to exist. Here, we provide a mathematical tool to help understand how and when cultural evolution operates more quickly than biological evolution. The speed of biological evolution can be quantified mathematically by equations based on diffusion (Kimura, 1980), and these have been applied to language evolution (e. g. De Boer et al., 2020). For cultural evolution the must be modified, because it can occur in social networks with a heavy-tailed neighborhood distribution: some individuals are disproportionally influential (Amaral et al., 2000; Onnela et al., 2007). This means that the ordinary diffusion equations used in biology become fractional diffusion equations (Metzler & Klafter, 2000)

2. Method and Result

In analogy to biological evolution, a fixed-size population (that does not change biologically) is modeled where two cultural variants compete. Conditional fixation time ϑ (the number of interactions for a variant to take over the population, when it does so) then follows the following fractional differential equation:

$$\begin{aligned} &\frac{c}{N^{\alpha}} \cdot \Big[(p^{\alpha}(1-p) + p(1-p)^{\alpha})\, {}_{RZ}D_x^{\alpha} \vartheta(p) \\ &-\tan\frac{\pi\alpha}{2} \cdot (p^{\alpha}(1-p) - p(1-p)^{\alpha})\, {}_{F}D_1^{\alpha} \vartheta(p) \Big] = -p \end{aligned} \tag{1}$$

with boundary conditions $\vartheta(0) = \vartheta(1) = 0$. The fixation time in absolute time (generations, not interactions) is then $\tau(p) = \vartheta(p)/(p \cdot N)$. N is the population size, p the proportion of the variant at the start, $1 < \alpha < 2$ is the parameter of the power law distribution that determines the heavy-tailedness of the neighborhood

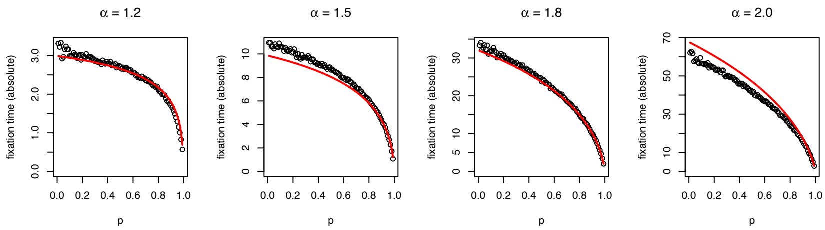


Figure 1. Fixation times (in arbitrary units of time) for culturally evolving populations with *N*=100 on social networks with differently distributed neighborhood sizes. Red lines represent solutions of the diffusion equation, and dots represent direct simulation. The leftmost graph corresponds to the most heavy-tailed distribution, while the rightmost graph corresponds to normally distributed neighborhood sizes (biological evolution). Vertical scales have different ranges to highlight the differences in curve shape.

distribution (the lower α, the more heavy-tailed) and c is a constant that depends on the precise shape of the neighborhood distribution. The operators ${}_{RZ}D_x^{\alpha}$ and ${}_{F}D_1^{\alpha}$ are *fractional derivatives* that generalize the second and first derivative, respectively (Herrmann, 2014, eqs 5.71 and 5.80). The first line of eq. 1 behaves similarly to ordinary diffusion (as in biological evolution). The second line introduces a drift away from the middle, which speeds up fixation.

Results of solving this equation and of directly simulating the evolutionary process are shown in Fig. 1. The correspondence between the model and the simulations is not perfect, but the diffusion model is a reasonable approximation. Importantly, fixation time is faster on the more heavy-tailed social networks, and because the curves are initially flatter, fixation time is less dependent on the initial prevalence of a culturally transmitted item in these cases.

3. Discussion and conclusion

Equation (1) has to the best of our knowledge not been described before, although similar systems have been studied (Carro et al., 2016). It allows us to estimate the time it takes for culture – formed by a population of agents interacting in a social network – to change; in the context of language evolution for instance for linguistic innovations to spread. The equation allows us to link properties of the social network (its size and the heavy-tailedness of its neighborhood distribution) and the initial frequency of a variant to the time it takes for this variant to spread. At the moment, the equation can only model drift, not selection (cultural variants do not differ in fitness), but the results from Fig. 1 suggest that even in this case cultural changes can spread rapidly, and their spread depends less strongly on their initial prevalence than in biological evolution. The form of the equation allows us to link it with the rich existing literature on fractional diffusion (Metzler & Klafter, 2000). Much remains to be done: extending the equation to differences in fitness, for instance, but also determining realistic values for α. Literature on *modern* social networks exists (Onnela et al., 2007), but it is an open question whether (pre-)historic cultural networks had the same structure.

Acknowledgements
This research received funding from the Flemish Government (AI Research Program).

References

Amaral, L. A. N., Scala, A., Barthélémy, M., & Stanley, H. E. (2000). Classes of small-world networks. *Proceedings of the National Academy of Sciences*, *97*(21), 11149–11152. https://doi.org/10.1073/pnas.200327197

Baronchelli, A., Chater, N., Christiansen, M. H., & Pastor-Satorras, R. (2013). Evolution in a changing environment. *PloS One*, *8*(1), e52742.

Carro, A., Toral, R., & San Miguel, M. (2016). The noisy voter model on complex networks. *Scientific Reports*, *6*(1), 24775.

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, *31*, 489–558.

De Boer, B., & Thompson, B. (2018). Biology-culture co-evolution in finite populations. *Scientific Reports*, *8*(1), 1209.

De Boer, B., Thompson, B., Ravignani, A., & Boeckx, C. (2020). Evolutionary Dynamics Do Not Motivate a Single-Mutant Theory of Human Language. *Scientific Reports*, *10*(451), 1–9. https://doi.org/10.1038/s41598-019-57235-8

Herrmann, R. (2014). *Fractional Calculus: An Introduction for Physicists* (2nd Edition). World Scientific.

Kimura, M. (1980). Average time until fixation of a mutant allele in a finite population under continued mutation pressure: Studies by analytical, numerical, and pseudo-sampling methods. *Proceedings of the National Academy of Sciences*, *77*(1), 522–526.

Kirby, S., & Hurford, J. R. (1997). Learning, Culture and Evolution in the Origin of Linguistic Constraints. In P. Husbands & I. Harvey (Eds.), *Fourth European Conference on Artificial Life* (pp. 493–502). MIT Press,.

Metzler, R., & Klafter, J. (2000). The random walk's guide to anomalous diffusion: A fractional dynamics approach. *Physics Reports*, *339*(1), 1–77.

Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., & Barabási, A.-L. (2007). Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, *104*(18), 7332–7336. https://doi.org/10.1073/pnas.0610245104

Self-domestication traits in vocal learning mammals

Koen de Reus[*1,2,3], Antonio Benítez-Burraco[4], Taylor A. Hersh[1,5], Nikita Groot[6], Megan L. Lambert[,7], Katie E. Slocombe[8], Sonja Vernes[6], and Limor Raviv[9,10]

[*]Corresponding Author: koen.dereus@mpi.nl

[1]Comparative Bioacoustics Group, MPI for Psycholinguistics, Nijmegen, NL

[2]Artificial Intelligence Laboratory, Vrije Universiteit Brussel, Brussels, BE

[3]Donders Centre for Cognition, Radboud University, Nijmegen, NL

[4]Department of Spanish, Linguistics & Theory of Literature, University of Seville, Seville, Spain

[5]Marine Mammal Institute, Oregon State University, Newport, USA

[6]School of Biology, University of St Andrews, St Andrews, UK

[7]Messerli Research Institute, University of Veterinary Medicine Vienna, Vienna, Austria

[8]Department of Psychology, University of York, UK

[9]LEADS group, MPI for Psycholinguistics, Nijmegen, NL

[10]cSCAN, University of Glasgow, UK

Both vocal learning and the human self-domestication hypothesis have been posited as relevant phenotypes for explaining language emergence in our species. *Vocal learning* (VL; i.e., the ability to learn new vocalizations or modify existing ones based on auditory experience) is a prerequisite for human speech acquisition and development, potentially providing insights into the *biological* underpinnings of language (Jarvis, 2019; Vernes et al., 2021). Until now, VL traits have been observed in groups of birds (parrots, songbirds, and hummingbirds) and mammals (humans, bats, elephants, cetaceans, and pinnipeds) (Petkov & Jarvis, 2012), with limited evidence found in non-human primates like marmosets (Takahashi et al., 2017). *Self-domestication* (SD; i.e., selective pressures against aggression and in favor of prosociality that give rise to a set of cognitive, behavioral, and physiological traits collectively known as the domestication syndrome) was recently invoked to potentially provide insights into language evolution through a *cultural* mechanism (Hare, 2017; Thomas & Kirby, 2018; Benítez-Burraco & Progovac, 2020; Raviv & Kirby, 2023). To date, SD has only been found in a narrow set of species (humans, bonobos, elephants, and perhaps marmosets; Hare, 2017; Ghazanfar et al., 2020; Raviv et al., 2023).

Both VL and SD are associated with two relevant traits that have been linked to language emergence. Specifically, despite variability in VL capacities (Vernes et al., 2021), vocal learners possess an improved vocal ability to share information with others (Nowicki & Searcy, 2014), helping them to better modulate social interactions. Similarly, despite variability in SD traits (Sánchez-Villagra et al., 2016), domesticated species show reduced aggression and increased prosocial behaviors, supporting more complex community ties (Burkart et al., 2018; Raviv et al., 2019; Dunbar, 1993). Interestingly, some domesticated species also show increased vocal complexity compared to their wild conspecifics, including Bengalese finches (Okanoya, 2017) and certain mammals (cats: Nicastro, 2004; dogs: Feddersen-Petersen, 2000; foxes: Gogoleva et al., 2011; cavies: Monticelli & Ades, 2011). This increase in vocal complexity may be due to altered stress responses as animals become tame, consequently leading to changes in dopaminergic activity in neural circuits crucial for VL (O'Rourke et al., 2021).

Could there be a link between these phenotypes? For example, do VL species also show a large number of SD traits? Given the potential link between the two phenotypes, we predict that some characteristic domestication traits, such as increased social tolerance, will be found across vocal learners. Testing to what extent these two phenotypes may overlap can improve our understanding of human language evolution, and help identify which non-human animal models are most useful for comparative language evolution studies.

Here, we conducted an exploratory cross-species comparison of SD traits in vocal learners. We focused this study on six VL mammals (elephants, bats, dolphins, whales, seals, and marmosets), of which only elephants have been the subject of previous SD research (Raviv et al., 2023). We looked at more than 20 behavioral and biological SD traits derived from previous work (e.g., Shilton et al, 2020). Besides elephants, our analysis did not reveal clear *morphological* SD traits in our studied species. For example, we did not observe a morphological reduction in the size of the skull, face, and jaw, which is typical to domesticated species, likely due to ecological differences related to feeding and habitat preferences (e.g., terrestrial vs. aquatic). Nevertheless, preliminary results show that the most crucial *behavioral* traits of SD (i.e., prosociality, exploratory behavior, and play) are shared across the VL mammals we investigated. This finding underscores the idea that, when taken together, these traits may be linked to the evolution of language, possibly through a shared mechanism. In future work, we plan to extend our comparisons to birds and include a control species.

Acknowledgements

TAH was supported by Independent Max Planck Research Group Leader funding to Andrea Ravignani. SCV was supported by a UKRI Future Leaders Fellowship, (MR/T021985/1) and ERC Consolidator Grant (101001702; BATSPEAK).

References

Benítez-Burraco, A., & Progovac, L. (2020). A four-stage model for language evolution under the effects of human self-domestication. *Language & Communication*, *73*, 1-17.

Burkart, J., Guerreiro Martins, E., Miss, F., & Zürcher, Y. (2018). From sharing food to sharing information: cooperative breeding and language evolution. *Interaction Studies*, *19*(1-2), 136-150.

Dunbar, R. I. (1993). Coevolution of neocortical size, group size and language in humans. *Behavioral and brain sciences*, *16*(4), 681-694.

Feddersen-Petersen, D. U. (2000). Vocalization of European wolves (Canis lupus lupus L.) and various dog breeds (Canis lupus f. fam.). *Archives Animal Breeding*, *43*(4), 387-398.

Ghazanfar, A. A., Kelly, L. M., Takahashi, D. Y., Winters, S., Terrett, R., & Higham, J. P. (2020). Domestication phenotype linked to vocal behavior in marmoset monkeys. *Current Biology*, 30(24), 5026-5032.

Gogoleva, S. S., Volodin, I. A., Volodina, E. V., Kharlamova, A. V., & Trut, L. N. (2011). Explosive vocal activity for attracting human attention is related to domestication in silver fox. *Behavioural Processes*, *86*(2), 216-221.

Hare, B. (2017). Survival of the friendliest: Homo sapiens evolved via selection for prosociality. *Annual Review of Psychology*, *68*, 155-186.

Jarvis, E. D. (2019). Evolution of vocal learning and spoken language. *Science*, *366*(6461), 50-54.

Monticelli, P. F., & Ades, C. (2011). Bioacoustics of domestication: alarm and courtship calls of wild and domestic cavies. *Bioacoustics*, *20*(2), 169-191.

Nicastro, N. (2004). Perceptual and Acoustic Evidence for Species-Level Differences in Meow Vocalizations by Domestic Cats (Felis catus} and African Wild Cats (Felis silvestris lybica). *Journal of Comparative Psychology*, *118*(3), 287.

Nowicki, S., & Searcy, W. A. (2014). The evolution of vocal learning. *Current Opinion in Neurobiology*, *28*, 48-53.

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.

O'Rourke, T., Martins, P. T., Asano, R., Tachibana, R. O., Okanoya, K., & Boeckx, C. (2021). Capturing the effects of domestication on vocal learning complexity. *Trends in Cognitive Sciences*, *25*(6), 462-474.

Petkov, C. I., & Jarvis, E. D. (2012). Birds, primates, and spoken language origins: behavioral phenotypes and neurobiological substrates. *Frontiers in Evolutionary Neuroscience*, *4*, 12.

Raviv, L., Meyer, A., & Lev-Ari, S. (2019). Larger communities create more systematic languages. *Proceedings of the Royal Society B*, *286*(1907), 20191262.

Raviv, L., Jacobson, S. L., Plotnik, J. M., Bowman, J., Lynch, V., & Benítez-Burraco, A. (2023). Elephants as an animal model for self-domestication. *Proceedings of the National Academy of Sciences*, *120*(15), e2208607120.

Raviv, L., & Kirby, S. (2023). Self-domestication and the Cultural Evolution of Language. In J.J. Tehrani, J. Kendal & R. Kendal (eds), *The Oxford Handbook of Cultural Evolution.*

Sánchez-Villagra, M. R., Geiger, M., & Schneider, R. A. (2016). The taming of the neural crest: a developmental perspective on the origins of morphological covariation in domesticated mammals. *Royal Society Open Science*, *3*(6), 160107.

Shilton, D., Breski, M., Dor, D., & Jablonka, E. (2020). Human social evolution: self-domestication or self-control?. *Frontiers in Psychology*, 134.

Takahashi, D. Y., Liao, D. A., & Ghazanfar, A. A. (2017). Vocal learning via social reinforcement by infant marmoset monkeys. *Current Biology*, *27*(12), 1844-1852.

Thomas, J., & Kirby, S. (2018). Self domestication and the evolution of language. *Biology & Philosophy*, *33*, 1-30.

Vernes, S. C., Kriengwatana, B. P., Beeck, V. C., Fischer, J., Tyack, P. L., Ten Cate, C., & Janik, V. M. (2021). The multi-dimensional nature of vocal learning. *Philosophical Transactions of the Royal Society B*, *376*(1836), 20200236.

3SG is the most conservative subject marker across languages: An exploratory study of rate of change

Peter Dekker*[1], Sonja Gipper[2], and Bart de Boer[1]

*Corresponding Author: peter.dekker@ai.vub.ac.be
[1]AI Lab, Vrije Universiteit Brussel, Brussels, Belgium
[2]Institut für Linguistik, Philosophische Fakultät, Universität zu Köln, Cologne, Germany

In this paper, we investigate the rate of change of different person-number subject markers. We perform a cross-linguistic study on the dissimilarity between proto and modern forms, showing that 3SG is the most conservative subject marker across languages. We discuss the mechanisms that could explain this diachronic pattern, such as frequency of use, markedness, and attractor lengths. Our exploratory analysis highlights how existing linguistic datasets can be used to study new research questions.

1. Introduction

Many languages mark the person and number of the subject by means of a bound morpheme on the verb (Siewierska, 2013; e.g., *walk-s*, with *-s* marking the third person singular). These verbal person-number subject markers are known to change over time, with certain diachronic changes in paradigms of subject markers being more probable than others (Cysouw, 2001); for instance, the form for 3SG is more likely to extend to other persons than vice versa (Baerman, 2005). But it is less well studied whether there is a difference in the *rate of change* across the different person-number combinations. Are certain subject markers more prone to change than others? This is the question we set out to investigate in this paper.

We perform an exploratory quantitative study of rates of change for six different person-number combinations – first, second, and third person, each in singular and plural – in a sample of 310 languages (Seržant & Moroz, 2022, data publication: Seržant, 2021). We find that 3SG is the most conservative subject marker across languages. We then discuss these findings in light of possible factors that may be responsible for this pattern. We suggest that, in line with previous work (Pagel, Atkinson, & Meade, 2007; Hoekstra & Versloot, 2019), our data hint at an important role for frequency in the rate of change of subject markers, as it could plausibly be the driving factor in the pattern we observe, while also relating to other possible explanations such as markedness and attractor lengths.

2. Method

Our sample of 310 modern languages associated with 15 proto-languages constitutes a subset of data[1] from Seržant (2021) who created a sample of subject markers in 383 languages from 53 families, as well as the reconstructed forms in their respective proto-languages, for six grammatical persons: 1SG, 2SG, 3SG, 1PL, 2PL, 3PL.[2] We calculate the Levenshtein distance (Heeringa, 2004) between proto and modern forms.[3] We use this degree of dissimilarity between proto and modern forms as a proxy for *rate of change*, i.e., amount of change over time period. Given the uncertainties regarding the estimation of the age of language families (Maurits, de Heer, Honkola, Dunn, & Vesakoski, 2020), our approach is agnostic with respect to the potentially different ages of the language families and proto-languages.[4]

The results of our distance calculation depend on the reconstructions of proto-forms, about which there is not always a consensus or which might represent an abstraction. Therefore, when comparing proto to modern forms, we do not assume that these comparisons necessarily represent concrete changes with historical reality. Rather, we aim to search for a general signal of cross-linguistic differences between subject markers. Moreover, reconstructed proto-forms generally give an underestimation of change, as traits of the proto-language not preserved in the daughter languages are not included in the reconstructed form (Campbell, 2013, p. 144). Despite these remaining uncertainties, we think this comparison between proto-forms and modern forms can serve as a fruitful first exploration of our research question, sketching an approach to explore an existing cross-linguistic dataset to find evidence for a novel linguistic question (cf. Ladd, Roberts, & Dediu, 2015).

We analyse the data using a mixed linear model (details in SI). The Levenshtein distance constitutes the response variable and person and number serve as

[1]All code of this paper can be found in `https://zenodo.org/doi/10.5281/zenodo.10722183` and the GitHub repository `https://github.com/peterdekker/changesubjectmarkers`. The Supplementary Information of this paper contains additional information on the technical details of the applied method.

[2]The dataset does not report forms that show a contrast in terms of clusivity, nor dual or paucal subject markers.

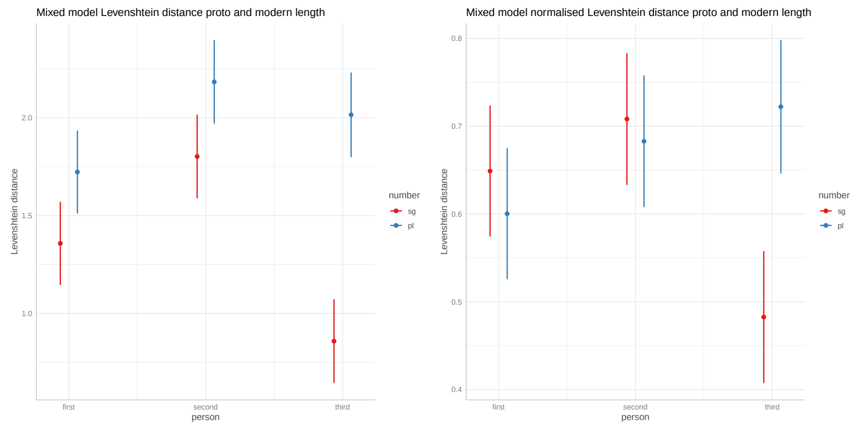
[3]For this exploratory analysis, we calculate the distance between orthographic forms as reported in the dataset. A more fine-grained analysis could be conducted in the future by using phonetic forms or even taking into account phonetic features (cf. List, 2012; Mortensen et al., 2016).

[4]Assuming that any specific age would apply in the same way to all person markers of a given language, we propose that family age can be neglected in our analysis. Also, Rama and Wichmann (2020, Table 6) show that family ages are in the same order of magnitude, for a sample overlapping ours. Moreover, in general the age of proto-languages is bounded by the time depth of reconstruction of the comparative method: maximum 6,000–10,000 years (Campbell, 2013, p. 341). For a more precise treatment of proto-language age, one could include a phylogenetic model in the analysis (e.g. Hahn & Xu, 2022).

predictors, with an interaction between person and number. We use clade as a random effect, because data points from languages in the same clade in a family should be treated as not fully independent, even more so because the Levenshtein distances are calculated with respect to the same proto-language. This random effect also partially addresses the potentially different ages of proto-languages. We report normalised and unnormalised Levenshtein distance. Unnormalised Levenshtein distance corresponds to a theory of a fixed rate of change *per form*: every timestep, there is a certain probability that 1 segment in the form will change. Whatever the length of the form, a change of 1 segment gives a Levenshtein distance of 1. On the other hand, normalised Levenshtein distance (distance divided by the length of the longest form), is based on a theory of a fixed rate of change *per phoneme* in a language. This assumes regular sound change, where a certain segment is substituted by another segment in all the forms in the language. For example, if in a language, the words ab and $abab$ have changed to ac and $acac$, due to the regular sound change $b \rightarrow c$, both receive a normalised distance 0.5, assigning the same score to forms affected by the same process of change. In this way, normalisation accounts for the fact that long forms have a higher chance of containing phonemes subject to regular sound change. Normalised Levenshtein distance is commonly used in phylogenetic reconstruction of language families (Serva & Petroni, 2008), which depends to a large extent on regular sound changes. For our purposes, to identify the rate of change per person marker, agnostic of the processes of change that are involved, we believe unnormalised Levenshtein distance is most suitable. However, we also report normalised Levenshtein distance for comparison.

3. Results

The predictions of the mixed linear model are shown in Figure 1. In the unnormalised model (Figure 1a), 3SG is the most conservative, while 2PL and 3PL are most innovative. Overall, singular forms are, on average, more conservative than their plural counterparts. The normalised (Figure 1b) model also shows 3SG as most conservative, while the difference between singular and plural can no longer be observed for first and second person. In sum, the most robust finding across both models is that 3SG is the most conservative among the six subject markers.



(a) Unnormalised Levenshtein distance (b) Normalised Levenshtein distance

Figure 1.: Predictions and 95% confidence intervals of a mixed linear model, with Levenshtein distance predicted from person and number (interaction), with clade as random effect. Higher values signify higher rates of change.

4. Discussion

We now turn to some possible explanations for our finding that in our analysis 3SG is the most conservative subject marker across languages. One factor that could arguably lead to this pattern is frequency of use, which has been shown to influence language change in at least two ways (Bybee & Thompson, 1997; Diessel, 2007, pp. 117–123; Hoekstra & Versloot, 2019): a conserving effect on morphology and a reducing effect on phonetics (Hinskens, 2011, p. 442). Both types of frequency effects are relevant for 3SG subject markers, as these tend to be both more conservative (our study) and shorter (Seržant & Moroz, 2022) than other subject markers.

Let us first turn to the conserving effect of frequency, based on the observation that high frequency of use reinforces the representation of a form, thereby preventing high-frequency irregular forms from becoming regularised (Diessel, 2007). Our finding that 3SG is the most conservative subject marker is consistent with this conserving effect of frequency, as there is evidence that in spoken language 3SG is the most frequent type of subject (e.g., Bybee, 1985, p. 71 on Spanish; Scheibman, 2001, p. 68, on American English; Seržant & Moroz, 2022, pp. 5–7, on Russian).

Regarding the second, reducing effect of frequency, it is also consistent with the rates of changes for the different persons presented in section 3: specifically the unnormalised model shows some parallels to the attractor lengths for the different persons reported in Seržant and Moroz (2022, Figure 2), which were calculated on the basis of the same dataset.[5] Seržant and Moroz (2022) attribute the different attractor lengths to the reducing effect of frequency, with 3SG having the shortest attractor length.

The most extreme case of this is zero marking, which is cross-linguistically more common for 3SG than for other persons (Cysouw, 2001, pp. 53–58; Bickel, Witzlack-Makarevich, Zakharko, & Iemmolo, 2015, pp. 47–48). Moreover, proto-forms reconstructed as zero seem to be relatively conservative in our dataset.[6] Again, this is consistent with the finding that 3SG zero is more common in some families than others, i.e. that it is to some extent a genealogical phenomenon (see summary in Cysouw, 2001, pp. 53–58). A possible explanation for this conservative behaviour of 3SG zero in particular, at least in some cases, could be that some linguistic systems depend on 3SG to be zero-marked, such as in omnipredicative languages where all open lexical classes are basically predicates (Launey, 2004) – see Cristofaro (2021) for further possible factors that may lead to the non-development of a marker for 3SG. So possibly, the conservative nature of 3SG zero forms in combination with the generally low potential for change due to its short attractor length could explain our results instead of or in addition to frequency, although these factors relate to frequency.

Another factor that may have an influence on the rate of change in person markers is markedness. In a feature-based description of subject markers, it is generally assumed that the first and second person are more marked than the third person, as the latter does not exhibit the features of being a speech act participant and of being the author of the utterance (Buchler & Freeze, 1966, p. 81; Buchler, 1967, p. 42; Nevins, 2007). Furthermore, plural is more marked than singular (Cysouw, 2007, p. 6), which results in the lowest markedness for 3SG. In general, frequency and markedness go hand in hand, with marked forms also being less frequent (Bybee, 2010). Baerman (2005) suggests that markedness may explain the cross-linguistic tendency for it to be more likely that other persons (notably 1/2SG) take over the form of 3SG than vice versa. As the least marked and hence 'default' form, 3SG is more likely to extend to other persons. This is consistent with our finding that 3SG is the most conservative marker, as in this scenario, 3SG remains unchanged.

There are further aspects that will be necessary to integrate in a full investigation of rate of change in subject markers. For instance, it is clear that social dynamics impact on the rate of change of linguistic items, such as community size (Nettle, 1999). In addition, Cristofaro (2021) emphasises that in diachronic typol-

[5]However, our results for rate of change are not just an artefact of the lengths of the markers, as the normalised model (Figure 1b), where length of the person markers has largely been removed as a factor, still partially follows the patterns of the attractor lengths from Seržant and Moroz (2022), at least for the singular forms.

[6]For 3SG, in 82 out of 132 cases where the proto-form is zero, we observe a modern form that is also zero (62%).

ogy, it is important to take the different diachronic paths into account that can lead to a typological pattern. For our research question, this means that we should not only look at the overall rate of change of person markers, but also at the different diachronic paths that lead to more conservative 3SG. Moreover, it is necessary to tease apart frequency effects on rate of change from those on typologically preferred patterns. Cathcart, Herce, and Bickel (2022) present a study that suggests that frequency, rather than impacting on the *rate of change*, has an influence on *long-term preferences*, where more frequent lemmata in Romance languages are more likely to exhibit a stem alternation than less frequent ones. However, no influence of lemma frequency on rate of change was observed.

Finally, the role of processing in the change of subject markers and in language change in general will be a promising avenue for future investigation (see Bambini et al., 2021). There is some pioneering research on the effect of markedness in person agreement on online processing. In an ERP experiment, Alemán Bañón and Rothman (2019) find that in agreement mismatches in Spanish, there is a stronger P600 effect[7] when a 1SG subject is used with a mismatching 3SG verb, than in cases where a 3SG subject is combined with a mismatching 1SG verb. Such findings are highly relevant for investigating cases where the form of one person marker extends to other persons, but also for tendencies regarding rate of change across different person markers in general. Integrating the different strands of evidence will be an intriguing topic for future research.

5. Conclusion

In this paper, we showed an exploratory approach of using an existing linguistic dataset for a new research question. We found that 3SG is the most conservative subject marker and argued that frequency of use and, relatedly, markedness seem to be important factors influencing the rate of change of person markers. We would furthermore like to highlight the correlation with proposed cross-linguistic attractor lengths of different persons and the presence of zero markers. Further research will be necessary to tease these factors apart.

Acknowledgements

We are grateful to the reviewers of this article and an earlier submission for useful feedback. An earlier version of this work was presented to historical linguists at the University of Cologne; thanks to the audience for their comments. We would like to thank Yannick Jadoul for help with integrating R code into Python. All remaining errors are ours. PD was supported by a PhD Fellowship fundamental

[7] In recent research, it is argued that the P600 represents the degree of difficulty of integrating the target word in the unfolding utterance meaning (Aurnhammer, Delogu, Schulz, Brouwer, & Crocker, 2021; Aurnhammer, Delogu, Brouwer, & Crocker, 2023). The stronger the violation of expectations, the higher the P600 effect.

research (11A2821N) of the Research Foundation — Flanders (FWO). SG was funded by the University of Cologne Excellent Research Support Programme, funding line FORUM, project *Conversational Priming in Language Change*, as well as funding line Cluster Development Program, project *Language Challenges*.

References

Alemán Bañón, J., & Rothman, J. (2019). Being a participant matters: Event-related potentials show that markedness modulates person agreement in Spanish. *Frontiers in Psychology*, *10*, 746.

Aurnhammer, C., Delogu, F., Brouwer, H., & Crocker, M. W. (2023). The P600 as a continuous index of integration effort. *Psychophysiology*, *60*(9), e14302.

Aurnhammer, C., Delogu, F., Schulz, M., Brouwer, H., & Crocker, M. W. (2021). Retrieval (N400) and integration (P600) in expectation-based comprehension. *PLOS ONE*, *16*(9), e0257430.

Baerman, M. (2005). Typology and the formal modelling of syncretism. In G. Booij & J. van Marle (Eds.), *Yearbook of Morphology 2004* (pp. 41–72). Dordrecht: Springer Netherlands.

Bambini, V., Canal, P., Breimaier, F., Meo, D., Pescarini, D., & Loporcaro, M. (2021). Capturing language change through EEG: Weaker P600 for a fading gender value in a southern Italo-Romance dialect. *Journal of Neurolinguistics*, *59*, 101004.

Bickel, B., Witzlack-Makarevich, A., Zakharko, T., & Iemmolo, G. (2015). Exploring diachronic universals of agreement: Alignment patterns and zero marking across person categories. In J. Fleischer, E. Rieken, & P. Widmer (Eds.), *Agreement from a diachronic perspective* (pp. 29–51). Berlin/Boston: De Gruyter.

Buchler, I. R. (1967). The analysis of pronominal systems: Nahuatl and Spanish. *Anthropological Linguistics*, *9*(5), 37–43.

Buchler, I. R., & Freeze, R. (1966). The distinctive features of pronominal systems. *Anthropological Linguistics*, *8*(8), 78–105.

Bybee, J., & Thompson, S. (1997). Three frequency effects in syntax. In M. L. Juge & J. L. Moxley (Eds.), *Proceedings of the Twenty-Third Annual Meeting of the Berkeley Linguistics Society: General Session and Parasession on Pragmatics and Grammatical Structure* (Vol. 23, pp. 378–388). Berkeley.

Bybee, J. L. (1985). *Morphology: A study of the relation between meaning and form.* Amsterdam/Philadelphia: John Benjamins.

Bybee, J. L. (2010). Markedness: Iconicity, economy, and frequency. In *The Oxford Handbook of Linguistic Typology* (pp. 131–147). Oxford University Press.

Campbell, L. (2013). *Historical linguistics.* Edinburgh: Edinburgh University Press.

Cathcart, C., Herce, B., & Bickel, B. (2022). Decoupling speed of change and long-term preference in language evolution: Insights from Romance verb stem alternations. In *Proceedings of the Joint Conference on Language Evolution (JCoLE)* (pp. 101–108). Kanazawa, Japan: JCoLE.

Cristofaro, S. (2021). Typological explanations in synchrony and diachrony: On the origins of third person zeroes in bound person paradigms. *Folia Linguistica Historica*, *42*(1), 25–48.

Cysouw, M. (2007). Building semantic maps: The case of person marking. In M. Miestamo & B. Wälchli (Eds.), *New challenges in typology: Broadening the horizons and redefining the foundations* (pp. 225–247). Berlin: Mouton de Gruyter.

Cysouw, M. A. (2001). *The paradigmatic structure of person marking.* Unpublished doctoral dissertation, Katholieke Univ. Nijmegen, Nijmegen.

Diessel, H. (2007). Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology*, *25*(2), 108–127.

Hahn, M., & 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.

Heeringa, W. J. (2004). *Measuring dialect pronunciation differences using Levenshtein distance.* Unpublished doctoral dissertation, Rijksuniv. Groningen.

Hinskens, F. (2011). Lexicon, phonology and phonetics. Or: Rule-based and usage-based approaches to phonological variation. In P. Siemund (Ed.), *Linguistic universals and language variation* (pp. 425–466). Berlin: De Gruyter Mouton.

Hoekstra, E., & Versloot, A. P. (2019). Factors promoting the retention of irregularity: On the interplay of salience, absolute frequency and proportional frequency in West Frisian plural morphology. *Morphology*, *29*(1), 31–50.

Ladd, D. R., Roberts, S. G., & Dediu, D. (2015). Correlational studies in typological and historical linguistics. *Annual Review of Linguistics*, *1*(1), 221–241.

Launey, M. (2004). The features of omnipredicativity in Classical Nahuatl. *STUF - Language Typology and Universals*, *57*, 49–69.

List, J.-M. (2012). LexStat: Automatic detection of cognates in multilingual wordlists. In M. Butt, S. Carpendale, G. Penn, J. Prokić, & M. Cysouw (Eds.), *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH* (pp. 117–125). Avignon, France: Association for Computational Linguistics.

Maurits, L., de Heer, M., Honkola, T., Dunn, M., & Vesakoski, O. (2020). Best practices in justifying calibrations for dating language families. *Journal of Language Evolution*, *5*(1), 17–38.

Mortensen, D. R., Littell, P., Bharadwaj, A., Goyal, K., Dyer, C., & Levin, L. S. (2016). PanPhon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical papers* (pp. 3475–3484).

Osaka, Japan: The COLING 2016 Organizing Committee.

Nettle, D. (1999). Is the rate of linguistic change constant? *Lingua*, *108*(2-3), 119–136.

Nevins, A. (2007). The representation of third person and its consequences for person-case effects. *Natural Language & Linguistic Theory*, *25*(2), 273–313.

Pagel, M., Atkinson, Q. D., & Meade, A. (2007). Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature*, *449*, 717–720.

Rama, T., & Wichmann, S. (2020). A test of Generalized Bayesian dating: A new linguistic dating method. *PLOS ONE*, *15*(8), e0236522.

Scheibman, J. (2001). Local patterns of subjectivity in person and verb type in American English conversation. In J. L. Bybee & P. J. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 61–89). Amsterdam/Philadeophia: John Benjamins.

Serva, M., & Petroni, F. (2008). Indo-European languages tree by Levenshtein distance. *EPL (Europhysics Letters)*, *81*(6), 68005.

Seržant, I. (2021). *Dataset for the paper "Universal attractors in language evolution provide evidence for the kinds of efficiency pressures involved"*. https://zenodo.org/records/7641119: Zenodo.

Seržant, I. A., & Moroz, G. (2022). Universal attractors in language evolution provide evidence for the kinds of efficiency pressures involved. *Humanities and Social Sciences Communications*, *9*(1), 58.

Siewierska, A. (2013). Verbal person marking. In M. S. Dryer & M. Haspelmath (Eds.), *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology.

The role of generalisation in an Adaptive Resonance Theory model of learning inflection classes

Peter Dekker[*1], Heikki Rasilo[1], and Bart de Boer[1]

[*]Corresponding Author: peter.dekker@ai.vub.ac.be
[1]AI Lab, Vrije Universiteit Brussel, Brussels, Belgium

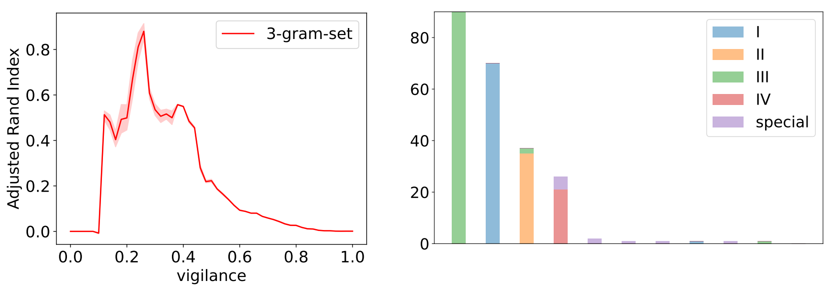
Humans are capable of generalising linguistic rules, e.g. by applying already acquired morphological patterns to unseen words (i.e. Prasada & Pinker, 1993; Krott, Baayen, & Schreuder, 2001). Inflection classes, groups of words that are inflected in the same way, help language users to deduce unseen word forms based on the patterns characteristic to the class (Milin, Filipović Durdević, & Moscoso Del Prado Martín, 2009; Veríssimo & Clahsen, 2014). Through this function in language processing, inflection classes can play a role in language change: inflection classes can attract new words to them (Round et al., 2022) and have been shown to become more distinct from one another over time (Enger, 2014). Any diachronic simulation of emergence or evolution of inflection classes needs a component for their acquisition on the individual level. In this study we investigate the role of generalisation in the individual learning task, with the ultimate goal of extending this to a diachronic model. We perform *unsupervised inflection class clustering* (cf. Guzmán Naranjo, 2020; LeFevre, Elsner, & Sims, 2021; Beniamine, Bonami, & Sagot, 2018 for related approaches) to investigate under which levels of generalisation a computer model is able to cluster verb paradigms together into inflection classes and which representations it learns. As a model, we use Adaptive Resonance Theory 1 (ART1) (Carpenter, 1987), a cognitively inspired neural network of category learning with one parameter, *vigilance*, controlling the degree of generalisation. The model learns in an online fashion, simulating the fact that a learner incrementally encounters data (Ackerman, Blevins, & Malouf, 2009; Blevins, Milin, & Ramscar, 2017). If the vigilance parameter is low, a new input sample is more likely to be added to an existing category, while if it is high, it is more likely that a new category will be created. The top-down weights in this two-layer network directly represent the features a certain category attends to, which provides interpretability of the learned representations (Grossberg, 2020)

We used the Latin present tense portion[1] of the Romance Verbal Inflection

[1]In Latin, inflection classes determine the inflection in the present tense and other tenses based on the present stem, but not in some other tenses like perfect (Pellegrini, 2019).

dataset (Beniamine, Maiden, & Round, 2020), which consists of phonetic forms of different paradigm cells for different verbs, as well as to which inflection classes these belong. We represent the data as trigrams, omitting temporal ordering of segments. As inputs to ART1 are binary vectors, we only register presence or absence of features. To combine the trigrams of all forms (1SG, 2SG, ... 3PL) for a verb (e.g. *stare* 'to stand') into one representation, we take the set of trigrams over the whole paradigm (i.e. presence of a trigram occurring in multiple forms is only registered once). 229 verbs (consisting of 971 trigram features) are run through the model two times. Figure 1a shows the classification accuracy for different vigilance values, evaluated using Adjusted Rand Index, a similarity measure between the inferred classification and the attested inflection classes. The model learns the inflection classes almost perfectly for a vigilance value of 0.25: this shows that a relatively high degree of generalisation (lower vigilance) is needed to obtain a good clustering. Analysis of the clustering of the best-performing model (Figure 1b) shows that the clusters roughly follow the real inflection classes in Latin, with the two first clusters perfectly matching with inflection classes III and I.

We conclude that ART1 is able to incrementally learn feature sets for groups of verb paradigms, that match well with known inflection classes for Latin. We find a narrow region of low vigilance parameter values (high generalisation) where the match is the best. An interesting next step would be to study evolution of inflection classes in an agent-based setting, where ART1 serves as an acquisition model for each agent. This setup would need an additional production model for transmitting word forms to other agents (cf. Hare & Elman, 1995; Cotterell, Kirov, Hulden, & Eisner, 2018; Parker, Reynolds, & Sims, 2018; Round et al., 2022 for agent-based models of inflection generation). If the agents would be initialised with word forms without a developed inflection class system, such experiments could also be used to study emergence of inflection classes.



(a) Clustering similarity to real inflection classes (Adjusted Rand Index), for different vigilance values, after 2 runs of data.

(b) Cluster analysis for vigilance 0.25 after 2 runs of data. Bar: discovered cluster, colour: real inflection class of assigned datapoints.

Figure 1.: Results ART1 on Latin present tense (trigram, set representation).

Acknowledgements

We would like to thank Sacha Beniamine, Lara Verheyen and the reviewers of this paper for useful feedback. PD was supported by a PhD Fellowship fundamental research (11A2821N) of the Research Foundation – Flanders (FWO). HR was funded by a Senior Postdoctoral Fellowship (1258822N) of the Research Foundation – Flanders (FWO).

References

Ackerman, F., Blevins, J. P., & Malouf, R. (2009). Parts and wholes: Patterns of relatedness in complex morphological systems and why they matter. *Analogy in grammar: Form and acquisition*, *54*, 82.

Beniamine, S., Bonami, O., & Sagot, B. (2018). Inferring Inflection Classes with Description Length. *Journal of Language Modelling*, *5*(3), 465.

Beniamine, S., Maiden, M., & Round, E. (2020). Opening the Romance Verbal Inflection Dataset 2.0: A CLDF lexicon. In *Proceedings of the 12th language resources and evaluation conference* (pp. 3027–3035).

Blevins, J. P., Milin, P., & Ramscar, M. (2017). The Zipfian Paradigm Cell Filling Problem. In F. Kiefer, J. Blevins, & H. Bartos (Eds.), *Perspectives on Morphological Organization* (pp. 139–158). BRILL.

Carpenter, G. A. (1987). A Massively Parallel Architecture for a Self-Organizing Neural Pattern Recognition Machine. 62.

Cotterell, R., Kirov, C., Hulden, M., & Eisner, J. (2018). On the Diachronic Stability of Irregularity in Inflectional Morphology. *arXiv:1804.08262 [cs]*.

Enger, H.-O. (2014). Reinforcement in inflection classes: Two cues may be better than one. *Word Structure*, *7*(2), 153–181.

Grossberg, S. (2020). A Path Toward Explainable AI and Autonomous Adaptive Intelligence: Deep Learning, Adaptive Resonance, and Models of Perception, Emotion, and Action. *Frontiers in Neurorobotics*, *14*.

Guzmán Naranjo, M. (2020). Analogy, complexity and predictability in the Russian nominal inflection system. *Morphology*, *30*(3), 219–262.

Hare, M., & Elman, J. L. (1995). Learning and morphological change. *Cognition*, *56*(1), 61–98.

Krott, A., Baayen, R. H., & Schreuder, R. (2001). Analogy in morphology: Modeling the choice of linking morphemes in Dutch. *Linguistics*, *39*(1).

LeFevre, G., Elsner, M., & Sims, A. D. (2021). Formalizing Inflectional Paradigm Shape with Information Theory.

Milin, P., Filipović Durdević, D., & Moscoso Del Prado Martín, F. (2009). The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from Serbian. *Journal of Memory and Language*, *60*(1), 50–64.

Parker, J., Reynolds, R., & Sims, A. D. (2018). A Bayesian Investigation of

Factors Shaping the Network Structure of Inflection Class Systems. *Proceedings of the Society for Computation in Linguistics*, 3.

Pellegrini, M. (2019). *Predictability in Latin inflection An entropy-based approach.* Unpublished doctoral dissertation, Università Degli Studi Di Bergamo.

Prasada, S., & Pinker, S. (1993). Generalisation of regular and irregular morphological patterns. *Language and Cognitive Processes*, *8*(1), 1–56.

Round, E., Mann, S., Beniamine, S., Lindsay-Smith, E., Esher, L., & Spike, M. (2022). COGNITION AND THE STABILITY OF EVOLVING COMPLEX MORPHOLOGY: AN AGENT-BASED MODEL. In *Joint Conference on Language Evolution.*

Veríssimo, J., & Clahsen, H. (2014). Variables and similarity in linguistic generalization: Evidence from inflectional classes in Portuguese. *Journal of Memory and Language*, *76*, 61–79.

Alignment vs conflict of interests in language evolution: Two pathways to high-level mindreading

Angelo D. Delliponti*[1], Sławomir Wacewicz[1], and Michael Pleyer[1]

*Corresponding Author: angelo.d.delliponti@gmail.com
[1]Center for Language Evolution Studies, Nicolaus Copernicus University in Toruń, Toruń, Poland

High-Level Mindreading (HLM), a type of controlled and reflective mentalizing activity, has been argued to be one of the core cognitive abilities underlying language (e.g.; Scott-Phillips, 2014; Wilson & Sperber, 2006). Most accounts of the evolutionary emergence of HLM in early humans put it in the context of cooperatively working towards common goals. Relatively less attention has been given to more competitive scenarios that assign greater prominence to the elements of rivalry between individuals (Tomasello, 2018; Witteveen, 2021). Here, we look at the latter aspect, arguing for its greater than currently appreciated relevance to the evolution of HLM. Specifically, we claim that:
- evolutionarily, HLM may derive not only from pressures on (i) *optimising communicative relevance* in the service of making cooperation effective, but also pressures on (ii) *epistemic vigilance* in the service of making cooperation stable;
- the relative importance of optimising relevance vs epistemic vigilance depends on a single predictive factor, i.e. the degree of alignment of interests between individuals. Highly aligned interests promote cooperating effectively, whereas a degree of conflict of interests promotes being epistemically vigilant so as to minimise the risk of deception and defection. In short, optimizing relevance improves coordination skills useful for cooperation, while epistemic vigilance creates a cognitive defense against attempts at deception.

It is widely agreed that the physical and social ecology of early humans involved a variety of contexts in which collaborative interactions were essential for one's fitness (such as big game hunting or cooperative breeding). Most accounts (Tomasello, 2018; Witteveen, 2021) highlight the benefits of cooperation and the need to evaluate the competence of the potential collaborators and the ways of efficiently coordinating joint action, which puts high demands on the cognitive skills related to the understanding of the mental states of others. This is indeed the case where the interests of group members are closely aligned. However, individuals in a group always tend to have partly conflicting interests, because they compete for the same limited resources, such as food, safety, or high-quality mates. Proportional to the degree of conflict of interest is the risk of defection and deception in communication, which in turn puts high demands on being able to accurately determine the trustworthiness and reliability of a potential partner.

We consider here the theoretical framework proposed by Heintz and Scott-Phillips (2023), in which relevance optimization, ostensive communication and epistemic vigilance (EV) played a crucial role in the evolution of human communication and language. Mentalizing abilities (i.e. theory of mind) are the evolutionary outcome of a process in which understanding others' communicative and informative intentions was an obligatory path to have open-ended, highly flexible and context-dependent, indefinitely recursive and voluntary communication (Scott-Phillips & Heintz, 2023).

We would like to complement this account by stating, as signalled above, that the degree of alignment of interests predicts two different evolutionary scenarios. In situations of high alignment of interests, i.e. mutualistic or near-mutualistic interactions where the risk and/or costs of partner defection are relatively small, it is more important to maximise the benefits of cooperative interaction. Such scenarios prioritise coordination and optimising relevance over being epistemically vigilant. Conversely, when the alignment of interests is smaller, and the risk of deception or partner defection is higher, it is more important to minimise those risks; this prioritises epistemic vigilance over relevance-optimisation. Importantly, the two paths are not mutually exclusive, and in fact both rest on the importance of cooperation, but bring to the forefront different aspects of cooperation subserved by different cognitive mechanisms: the effectiveness of cooperation (aided by relevance optimisation) vs the stability of cooperation (aided by epistemic vigilance).

What is the possible contribution of these two aspects of cognition to the origin of human communication? In a scenario without relevance optimization, there would not have been a bias that would have led to the origin of increasingly sophisticated communicative interactions; in a scenario without epistemic vigilance, the tools to deal with attempts at deception and manipulation made possible by a greater push for communication would have been lacking: language would never have appeared.

References

Heintz, C., & Scott-Phillips, T. (2023). Expression unleashed: The evolutionary and cognitive foundations of human communication. *Behavioral and brain sciences*, *46*, e1.

Scott-Phillips, T. (2014). *Speaking our minds: Why human communication is different, and how language evolved to make it special*. Bloomsbury Publishing.

Scott-Phillips, T., & Heintz, C. (2023). Animal communication in linguistic and cognitive perspective. *Annual Review of Linguistics, 9*, 93-111.

Tomasello, M. (2018). *A natural history of human thinking*. Harvard University Press.

Wilson, D., & Sperber, D. (2006). Relevance theory. *The handbook of pragmatics*, 606-632.

Witteveen, J. (2021). Biological markets, cooperation, and the evolution of morality. *The British Journal for the Philosophy of Science*.

The cognitive requirements for developing a multimodal communication system: Evidence from experimental semiotics and comparative cognition

Angelo Damiano Delliponti[1], Elizabeth Qing Zhang[2], Yen Ying Ng[1], and Michael Pleyer[1*]

[*]Corresponding Author: pleyer@umk.pl

[1]Center for Language Evolution Studies, Nicolaus Copernicus University in Toruń, Toruń, Poland

[2]School of Linguistic Sciences and Arts, Jiangsu Normal University, Xuzhou, China

Specifying the cognitive requirements for developing a structured, symbolic communication system is one of the most central tasks for accounts of what made humans 'language-ready' (Arbib 2012) and enabled them to evolve language. From an evolutionary perspective, it is also a central question to what degree these requirements are shared with other animals and how they evolved. One approach that sheds light on the processes and necessary requirements for the emergence of a symbolic communication system in interaction is that of experimental semiotics, the study of "novel forms of communication which people develop when they cannot use pre-established communication" (Galantucci et al., 2012).

In experimental semiotics, participants have to bootstrap communicative signals and establish a relation between a novel sign and its interpretation. In different paradigms, participants use different signals in different modalities to communicate meanings. For example, they can be asked to communicate via drawings, novel gestures, novel vocalisations, symbols, pantomime, or combining channels of different modalities (see, e.g. Nölle & Galantucci, 2022; for a review). What these experiments show is that participants are able to converge on a shared symbolic communication system, which over time also becomes increasingly structured.

Here, we adopt an evolutionary perspective on the cognitive requirements required for the establishment of shared symbolic systems in experimental semiotics paradigms. Specifically, we ask a) what are the cognitive requirements needed to explain the successful behaviour of participants in experimental semiotics studies; b) what are the evolutionary foundations and the possible evolutionary trajectories of these cognitive abilities.

In order to elucidate the first question, we make use of an existing database of experimental semiotics studies (Delliponti et al. 2023), and add to this database by adding a meta-analysis of the cognitive capacities needed for particular tasks that are explicitly mentioned in these studies. We analyzed the frequency of the cognitive abilities mentioned in the studies in the database created by Delliponti et al. (2023), standardized the labels, and additionally assigned cognitive abilities to general types like general cognition, social cognition, and motor cognition. Our analysis shows that although there is a wide variety of factors discussed in the 59 studies that were surveyed, some abilities occur more frequently, such as theory of mind, categorical perception, and memory factors; the same goes for cognitive types such as social cognition and general cognition.[1] Using such a meta-analytic approach therefore enables us to create a list of some of the most important abilities required for establishing a shared symbolic communication system.

To investigate the question as to the evolutionary foundation of these abilities, we review which of the specified necessary cognitive requirements are present in non-human animals, and if so, to which degree. For instance, regarding Theory of Mind (ToM), we know that human beings resort to metarepresentations, whereby they adopt second-order beliefs in order to anticipate other people's behavior. While many aspects of ToM seem to be shared with other animals, there also seem to be important differences (e.g. Call & Tomasello 2008; Beetle & Rosati 2021). For example there is evidence that chimpanzees use a type of simulative rather than metarepresentational ToM, in order to predict other agents' behavior (Lurz et al. 2022). In the case of Categorical Perception (CP), it was found in nonhuman animals across modalities, as in the case of the CP of sound or color. Field crickets (Wyttenbach et al., 1996), rodents (Sinnott & Mosteller, 2001), and macaques (Sandell, Gross & Bornstein, 1979), are among the cases of non-human animals with CP of sound and/or color, suggesting a deep evolutionary continuity. By adding insights from comparative cognition to the list created by the meta-analysis, we thereby can gather information not only on important abilities required for the establishment of a shared communication system, but also on the degree to which they are shared with other animals, and which aspects are potentially uniquely human. Overall, then, using an approach that combines insights from experimental semiotics and comparative cognition promises to shed light on the evolution of the cognitive requirements for the emergence of symbolic communication systems, and language more generally.

Acknowledgements

This research is part of the project No. 2021/43/P/HS2/02729 co-funded by the National Science Centre and the European Union's Horizon 2020 research and

[1] https://osf.io/w5t6p/?view_only=5b4ddaa01dea49d8b83f71b39a4281b4

innovation programme under the Marie Skłodowska-Curie grant agreement no. 945339.

References

Arbib, M. (2012). *How the Brain Got Language: The Mirror System Hypothesis*. Oxford University Press.

Bettle, R., & Rosati, A. G. (2021). The primate origins of human social cognition. *Language Learning and Development*, 17(2), 96-127.

Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? Thirty years later. *Trends in Cognitive Sciences*, 12, 187-192.

Delliponti, A., Raia, R., Sanguedolce, G., Gutowski, A., Pleyer, M., Sibierska, M., Placiński, M., Żywiczyński, P. & Wacewicz, S. (2023). Experimental Semiotics: A Systematic Categorization of Experimental Studies on the Bootstrapping of Communication Systems. *Biosemiotics*, *16*, 291-310.

Galantucci, B., Garrod, S., & Roberts, G. (2012). Experimental semiotics. *Language and Linguistics Compass*, 6(8), 477-493.

Lurz, R. W., Krachun, C., Mareno, M. C., & Hopkins, W. D. (2022). Do chimpanzees predict others' behavior by simulating their beliefs. *Animal Behavior and Cognition*, *9*(2), 153-175.

Nölle, J. & Galantucci, B. (2022). Experimental semiotics: past, present and future. In A. M. García & A. Ibánez (Eds.), *Routledge handbook of semiotics and the brain*. Routledge.

Sandell, J. H., Gross, C. G., & Bornstein, M. H. (1979). Color categories in macaques. *Journal of comparative and physiological psychology*, *93*(4), 626.

Sinnott, J. M., & Mosteller, K. W. (2001). A comparative assessment of speech sound discrimination in the Mongolian gerbil. *The Journal of the Acoustical Society of America*, *110*(4), 1729-1732.

Wyttenbach, R. A., May, M. L., & Hoy, R. R. (1996). Categorical perception of sound frequency by crickets. *Science*, *273*(5281), 1542-1544.

Defining the building blocks of pragmatic competence; the social context of language evolution

Miranda Dickerman[*1], Joseph Mine[2], Simon Townsend[2], Sabine Stoll[1]

[*]Corresponding Author: miranda.dickerman@uzh.ch
[1]Department of Comparative Language Science, University of Zurich, Zurich, Switzerland
[2]Department of Evolutionary Anthropology, University of Zurich, Zurich, Switzerland

Human communication relies on integrating signals with contextually available information, a process known as pragmatic competence. Pragmatic competence is neither fully unique to humans nor exclusive to language use (e.g., Arnold & Zuberbuhler 2013). This suggests that language evolved alongside an existing ability to integrate signals and context. In this study, we propose a framework of pragmatic competence and its evolution from a linguistic, psychological, and biological perspective. We aim to delineate the cognitive capacities which underlie pragmaticcompetence. To understand how these capacities co-evolved, we examine their presence in our closest relatives: the great apes.
We first introduce (a) a typology of information sources, which can be comparatively applied across species. This typology encompasses both signals emitted during communicative acts, (such as gestures, facial expressions, and linguistic signals), and information accessible outside of the communication process (such as knowledge pertaining to the environment or to the state of mind of the signaller). We then establish that, to a large extent, physical information sources are not comparable across species; rather, comparative relevance lies in the ability to access and interpret information sources.
Access to and interpretation of information sources, in turn, relies on different underlying capacities enabling derivation of and reasoning about information from distinct sources. From this point we expand on previous work on the evolutionary origins of pragmatic competence by establishing (b) a broad overview of the necessary cognitive capacities for accessing these information sources. We focus on the mechanisms that are minimally necessary for pragmatic competence. The most prominent of these mechanisms is theory of mind (e.g., Heintz & Phillips, 2023; Bar-On, 2021). We additionally identify and integrate

signal flexibility, or optionality (Watson et al., 2022), and predictive processing literature into the story of the capacities enabling pragmatic competence.

Finally, in (c), we apply the comparative method, examining evidence for the capacities underlying pragmatic competence in our closest relatives, in the domains of signal production and perception. To this end we draw on literature from diverse fields, including from work on primate communication (e.g., Wilke et al., 2017), predictive pre-processing of communicative input (e.g., Heilbron et al. 2022), and other comparative work on these cognitive capacities (e.g., Krupenye & Call, 2019). We find that the roots of the core components underlying pragmatic competence are present in our ape relatives, suggesting their presence in our lineage to be phylogenetically old. Deconstructing the components of pragmatic competence ultimately allows us to better disentangle the evolutionary trajectory of pragmatics, offering insight into the conditions under which language emerged.

Acknowledgements

This work was supported by the NCCR Evolving Language, Swiss National Science Foundation (no. 51NF40_180888).

References

Arnold, K., & Zuberbühler, K. (2013). Female putty-nosed monkeys use experimentally altered contextual information to disambiguate the cause of male alarm calls. *PLoS One*, *8*(6), e65660.

Bar-On, D. (2021). How to do things with nonwords: pragmatics, biosemantics, and origins of language in animal communication. *Biology & Philosophy*, *36*(6), 50.

Heilbron, M., Armeni, K., Schoffelen, J. M., Hagoort, P., & De Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, *119*(32), e2201968119.

Heintz, C., & Scott-Phillips, T. (2023). Expression unleashed: The evolutionary and cognitive foundations of human communication. *Behavioral and brain sciences*, *46*, e1.

Krupenye, C., & Call, J. (2019). Theory of mind in animals: Current and future directions. *Wiley Interdisciplinary Reviews: Cognitive Science*, *10*(6), e1503.

Watson, S. K., Filippi, P., Gasparri, L., Falk, N., Tamer, N., Widmer, P., ... & Glock, H. J. (2022). Optionality in animal communication: a novel framework for examining the evolution of arbitrariness. *Biological Reviews*, *97*(6), 2057-2075.

Wilke, C., Kavanagh, E., Donnellan, E., Waller, B. M., Machanda, Z. P., & Slocombe, K. E. (2017). Production of and responses to unimodal and multimodal signals in wild chimpanzees, Pan troglodytes schweinfurthii. *Animal Behaviour*, *123*, 305-316.

Balancing regularization and variation: The roles of priming and motivatedness

Loïs Dona[*1,2] and Marieke Schouwstra[3]

[*]Corresponding Author: lois.dona@mpi.nl
[1]Multimodal Language Department, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
[2]LEADS group, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
[3]Faculty of Humanities, University of Amsterdam, Amsterdam, The Netherlands

Languages exhibit both variation and regularities across different domains (Dryer & Haspelmath, 2013). A key question in language evolution research is how conventions that exist today arose from early stages of language with motivated variation. Priming, a tendency to repeat forms that you have seen before, is a potential mechanism driving regularization (Fehér, Wonnacott, & Smith, 2016; Pickering & Garrod, 2017; Schouwstra, Smith, & Kirby, 2020). On the other hand, research has shown that the use of motivated forms allows variation to persist (Mudd, Vos, & De Boer, 2022). In this work, we use an agent-based model to show when priming is able to cause regularization in a population, and when a level of variation is maintained. Maintenance of variation, even when there are dominant patterns, has been demonstrated in existing languages on lexical as well as structural levels (Napoli & Sutton-Spence, 2014; Flaherty, Schouwstra, & Goldin-Meadow, 2018; Napoli, Spence, & Quadros, 2017; Lutzenberger, Mudd, Stamp, Schembri, & Schembri, 2023).

In our models we focus on word order, in particular in connection with the balance between motivatedness and regularity. Traditionally, the regular nature of the ordering of Subject, Object and Verb in existing languages has been emphasised (Dryer, 2013). At the same time, improvised silent gesture experiments show that semantically motivated orders are preferred when there are no linguistic conventions yet: SOV when communicating about extensional events (the direct object is specific and concrete: e.g. boy kicks ball) and SVO when communicating about intensional events (the direct object is more abstract, and possibly dependent on the verb: e.g. boy thinks of ball) (Schouwstra & Swart, 2014; Motamedi, Wolters, Naegeli, Kirby, & Schouwstra, 2022). While existing languages do not generally exhibit the same level of variation as silent gesture, it was recently recognized that many languages exhibit some variation in word order, which therefore might best be analyzed as a gradient phenomenon (Levshina et al., 2023).

Here we model how syntactic priming affects the emergence of word order regularity in a population, when it exists in direct competition with motivatedness. Our simulations consists of N agents, who engage in communication for M steps. Every step, each agent produces a word order (which will be observed by another randomly chosen agent) according to a linear weighted combination of word orders that were observed before (priming) and word order preferences that are driven by the intensionality or extensionality of the event (motivatedness). In all simulations, production is fully based on motivatedness when there are no observations yet. As a consequence, there is more semantically motivated variation at the start of a simulation, which makes way for more regularity as agents align.

We executed simulations (all with 50 agents and 300 time steps) with different constant relative influences of motivatedness and priming and measure the proportion of the majority word order at each step. Results are presented in figure 1. Larger values of P(majority order) correspond to more regularity.

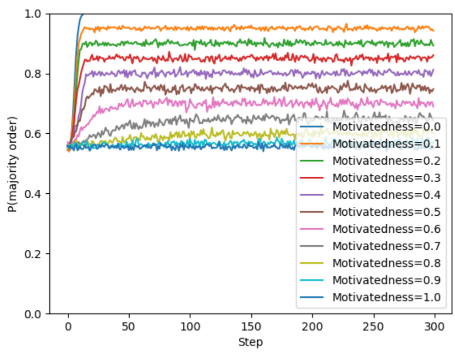


Figure 1.: Mean (over 50 runs) P(majority order) over time in simulations with different relative influences of motivatedness and priming

Figure 1 shows a direct relationship between the relative influence of priming and motivatedness on word order regularization. We see stronger reductions in variation when motivatedness has less influence than priming, but a small motivatedness bias will prevent the system from regularizing fully. This brief phase of stronger regularization, followed by a relatively stable situation with residual variation, is consistent with experimental data (Motamedi et al., 2022; Schouwstra, Naegeli, & Kirby, 2022) and with data from existing sign languages (Napoli & Sutton-Spence, 2014; Flaherty et al., 2018; Napoli et al., 2017). This reinforces the view that word order is a gradient phenomenon (Levshina et al., 2023). Our model shows that motivatedness and priming can jointly balance variation in the domain of word order, and provides a framework for other domains.

We extended these simulations to investigate the dynamics between priming and motivatedness under various social network structures (Lupyan & Dale, 2010; Richie, Yang, & Coppola, 2014; Lou-Magnuson & Onnis, 2018; Raviv, Meyer, & Lev-Ari, 2020). The average shortest path distance of a network influenced the extent of regularisation: regular networks retained the most variation, fully connected networks retained the least variation and others are in between. Importantly, in all network structures that were not fully connected, even a very small influence of motivatedness prevented full regularization of the system.

References

Dryer, M. S. (2013). Order of subject, object and verb. In M. S. Dryer & M. Haspelmath (Eds.), *The world atlas of language structures online.* Leipzig: Max Planck Institute for Evolutionary Anthropology.

Dryer, M. S., & Haspelmath, M. (Eds.). (2013). *Wals online (v2020.3).* Zenodo.

Fehér, O., Wonnacott, E., & Smith, K. (2016). Structural priming in artificial languages and the regularisation of unpredictable variation. *Journal of Memory and Language*, *91*, 158–180.

Flaherty, M., Schouwstra, M., & Goldin-Meadow, S. (2018). Do we see word order patterns from silent gesture studies in a new natural language. In *The evolution of language: Proceedings of the 12th international conference (evolangxii). ncu press. retrieved from http://evolang.org/torun/proceedings/papertemplate. html* (pp. 3991–1).

Levshina, N., Namboodiripad, S., Allassonnière-Tang, M., Kramer, M., Talamo, L., Verkerk, A., Wilmoth, S., Rodriguez, G. G., Gupton, T. M., Kidd, E., et al.. (2023). Why we need a gradient approach to word order. *Linguistics*(0).

Lou-Magnuson, M., & Onnis, L. (2018). Social network limits language complexity. *Cognitive Science*, *42*(8), 2790–2817.

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PloS one*, *5*(1), e8559.

Lutzenberger, H., Mudd, K., Stamp, R., Schembri, A. C., & Schembri, A. (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., Wolters, L., Naegeli, D., Kirby, S., & Schouwstra, M. (2022). From improvisation to learning: How naturalness and systematicity shape language evolution. *Cognition*, *228*, 105206.

Mudd, K., Vos, C. de, & De Boer, B. (2022). Shared context facilitates lexical variation in sign language emergence. *Languages*, *7*(1), 31.

Napoli, D. J., Spence, R. S., & Quadros, R. M. de. (2017). Influence of predicate sense on word order in sign languages: Intensional and extensional verbs. *Language*, *93*(3), 641–670.

Napoli, D. J., & Sutton-Spence, R. (2014). Order of the major constituents in sign languages: Implications for all language. *Frontiers in Psychology*, *5*, 376.

Pickering, M. J., & Garrod, S. (2017). Priming and language change. *The changing English language: Psycholinguistic perspectives*, 173–90.

Raviv, L., Meyer, A., & Lev-Ari, S. (2020). The role of social network structure in the emergence of linguistic structure. *Cognitive Science*, *44*(8), e12876.

Richie, R., Yang, C., & Coppola, M. (2014). Modeling the emergence of lexicons in homesign systems. *Topics in cognitive science*, *6*(1), 183–195.

Schouwstra, M., Naegeli, D., & Kirby, S. (2022). Investigating word order emergence: Constraints from cognition and communication. *Frontiers in Psychology*, *13*, 805144.

Schouwstra, M., Smith, K., & Kirby, S. (2020). The emergence of word order conventions: improvisation, interaction and transmission.

Schouwstra, M., & Swart, H. de. (2014). The semantic origins of word order. *Cognition*, *131*(3), 431–436.

Differences in distributional structure can lead to differences in similarity biases

Yuguang Duan[*1], Khuyen Le[2], and Gary Lupyan[1]

[*]Corresponding Author: yduan38@wisc.edu
[1]Department of Psychology, University of Wisconsin-Madison, Madison, USA
[2]Department of Psychology, University of California, San Diego, San Diego, USA

Some people think a beaver is more similar to a chimp than a turtle, while others think the opposite. What are the consequences of such differences in concepts/word-meanings on communication? And where might these differences come from? We conducted an experiment investigating whether exposure to different word co-occurrence patterns affects people's biases to rely on more taxonomic or thematic relations. English-speaking participants were asked to learn novel words (pseudo-words) in different co-occurring contexts (taxonomic, thematic, and neutral), and their similarity biases were measured over the learning process for each word group and across groups. Context exposure increases similarity biases for the matching context. Learned similarity biases persisted to novel lexical items. Overall, our findings show a causal link between being exposed to different distributional data and people's subsequent similarity ratings, providing a possible mechanism behind previously observed cross-linguistic differences in similarity biases.

1. Introduction

It is generally assumed that for a language to function as an effective communication system, both the forms and meanings must be closely aligned among speakers (e.g., Hutchins & Hazlehurst, 2006). However, different language users may have somewhat different word-meanings. These differences can be partially revealed by comparing people's similarity judgments, e.g., some people think a beaver to be more similar to a chimp than a turtle–prioritizing the common biological taxonomy of beaver and chimp. Others indicate that a beaver is more similar to a turtle, emphasizing the thematic relationship, in this case presumably the strong association with aquatic habitats (Martí, Wu, Piantadosi, & Kidd, 2023; Duan & Lupyan, 2023). These differences in some cases lead to communication misalignment (Duan & Lupyan, 2023). These findings raise a number of interesting questions such as where these differences come from, how languages adapt to tolerate them, and to what extent people's word-meanings (and even conceptual structure) diverge while maintaining communicative success? Here, we focus on the first question.

There is a long history of studying people's reliance on thematic and taxonomic factors in judging similarity relationships (e.g. Markman & Hutchinson, 1984; Lin & Murphy, 2001). One commonly observed trend is an increase then decrease in bias for taxonomic similarity with age (Smiley & Brown, 1979; Borghi & Caramelli, 2003; Brooks, Seiger-Gardner, & Sailor, 2014). Additionally, individual differences in similarity biases also vary with education (Luriia, 1976), language skills (Nation & Snowling, 1999), and the items being judged (Whitmore, Shore, & Smith, 2004). Another contributor is cultural and linguistic differences. People from Eastern cultures show more thematic bias compared to those from Western cultures (Nisbett & Masuda, 2003).

However, the mechanism through which these factors lead to individual differences in similarity biases is less understood. In a recent study, Le, Gao, Frank, and Carstensen (2023) point out that cross-cultural differences in similarity biases (such as those reported in Ji, Zhang, & Nisbett, 2004) may be explained by different statistical patterns in languages spoken by these people. They found that word similarity inferred by language models trained on different languages correlated with human similarity biases from the corresponding language. However, the correlational nature of this study fails to elucidate the extent to which language, as opposed to culture, contributes to human similarity biases. Additionally, the mechanism through which cross-cultural differences in language statistics produce differences in human similarity biases has not been investigated.

Our study addresses these gaps by using a word-learning experiment in which we manipulate linguistic patterns in which pseudo-words are embedded to examine the resulting patterns of similarity judgments. Following the results in Le et al. (2023), we hypothesize that individuals who are exposed to different word co-occurrence patterns will form different similarity biases. Specifically, our research questions are: (Q1) Do different inputs of word co-occurrence patterns (taxonomic, thematic, neutral) result in changes in the corresponding similarity biases? (Q2) Do changes in similarity biases caused by exposure to language patterns generalize to novel words?

2. Methods

2.1. *Materials*

The study focused on second-order word co-occurrences[1], which has been suggested to be the main source of model simulation of human similarity ratings (Paridon, Liu, & Lupyan, 2021). We generated three groups of pseudo-words consisting of a target word and three co-occurring words in taxonomic, thematic and neutral contexts, and five pairs of second-order co-occurring sentences for

[1]Occurrences in the same contexts. Consider 2 sentences: ”A chicken looks like a duck” and ”A goose looks like a duck”. ”Chicken” and ”goose” is an example of second-order co-occurrence because they both occur in the context of ”looks like a duck”, even if they do not co-occur within the same sentence.

each context. Taxonomic contexts focus on attributes and categories[2], thematic contexts focus on spatial and temporal occurrences[3], while neutral contexts are contexts that do not fall under taxonomic or thematic, serving as a baseline[4]. Nouns in the contexts were replaced with pseudo-words to minimize reliance on prior knowledge. All pseudo-words were generated by ChatGPT, and manually checked by two English speakers to ensure they were not in English dictionaries and had no phonetic similarity to other pseudo-words in the set.

2.2. *Participants*

We recruited 90 participants from Mechanical Turk, assigning them evenly to a taxonomic, neutral, or thematic condition. All participants indicated that their first language was English and that they resided in the US. The average age of participants is 40.2 (SD=11.7) and the average education level was 4.1 (SD=1.37)[5], with no significant differences among conditions.

2.3. *Procedure*

Participants were randomly assigned to one of three conditions: taxonomic, thematic, and neutral. In all conditions, participants completed a practice set of trials that familiarized them with the study, and three experimental sets of trials each of which consisted of a *prior* trial block and a *critical* trial block (see Figure 1).

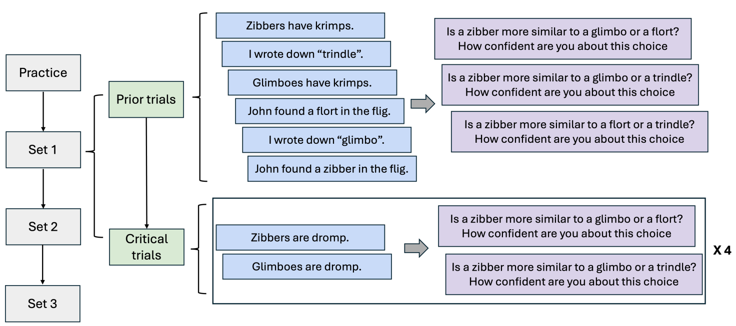


Figure 1. Procedure schematic for the taxonomic condition

Participants were asked to make a series of similarity judgments in which they

[2]For e.g., Zibbers have krimps. / Glimboes have krimps.
[3]For e.g., John found a zibber in the flig. / John found a flort in the flig.
[4]For e.g., I wrote down ”zibber”. / I wrote down ”trindle”.
[5]1 = Less than high school, 2 = High school diploma, 3 = Some college, no degree, 4 = 2-year/associate’s degree, 5 = Bachelor’s degree, 6 = Master’s degree, 7 = PhD, law, or medical degree.

chose which of two wordswas more similar to the target word, and then indicated their confidence on a 10-point Likert scale.

Prior trial block. Participants read three pairs of sentences (one in each context type). They then completed three similarity judgments (as described above) pitting pairs of co-occurring words against one another (taxonomic-thematic, taxonomic-neutral, and thematic-neutral).

Critical trial block. Participants saw four pairs of co-occurrence sentences corresponding to their assigned condition (e.g., in the taxonomic condition they only saw sentences involving the target word and the taxonomic match). The order of sentences was randomized, which could result in a delay between pairs of co-occurring sentences. After viewing each co-occurring sentence, they completed two similarity judgments pitting the co-occurring word they had just seen with the other co-occurring choices (e.g., in the taxonomic condition, the judgments pit the taxonomic match against the neutral and thematic match respectively).

After reading each sentence, participants saw a multiple-choice question that served as an attention check[6]. Participants who failed more than two attention checks were excluded from analysis. Participants' age and education level were collected at the end of the experiment for analysis. Pseudo-words, sentence order and set order were randomized across participants.

2.4. *Analytic Procedure*

We analyzed our data using two linear mixed-effects models. To answer Q1, we examined the effect of sequence order (order of four co-occurrences in each block) for each experimental block, age and education (all scaled), as well as their interactions with different types of comparison (neutral against taxonomic as the reference level). We used the following 'lmer' syntax: `bias ~ (sequence order + age + education) * Comparison Type + (1 | participant ID)`. Sequence order refers to one prior trial plus four critical trials within each set ($1 \sim 5$). There were six comparison types in total: for each condition, we used two comparison types – the similarity of the target word to the co-occurring word in this conditional type, compared to its similarity to the other two co-occurring words in prior trials. For example, in the taxonomic condition, we compared taxonomic against thematic, and taxonomic against neutral. Participants' biases were coded as positive if they preferred the condition-aligning choice and negative otherwise. We used their confidence rating for the absolute value of their bias. To answer Q2, we focused on how people's prior bias (bias for new pseudo-words) changed over the course of three sets. We therefore only considered the prior trial in each set. The model was `bias ~ (set order + age + education) * Comparison Type + (1 | participant ID)`.

[6]For e.g., after reading "John found a zibber in the flig", participants needed to choose whether the zibber was found "in the flig", "on the flig", or "under the flig".

3. Results

Clear evidence of learning is indicated by the main effect of sequence order ($b=1.59$, $p<0.005$), which means when people saw more pairs of sentences containing co-occurrence of two words in the corresponding type of context, their similarity bias for that type increased (Table 1 and Figure 2). While this learning slope didn't differ between different conditions, overall, people had a stronger taxonomic bias than thematic and neutral bias, and a stronger thematic than neutral bias (type[taxonomic against neutral]: $b=4.04$, $p<0.005$; type[taxonomic against thematic]: $b=3.81$, $p<0.005$; type[thematic against neutral]: $b=2.51$, $p<0.01$).

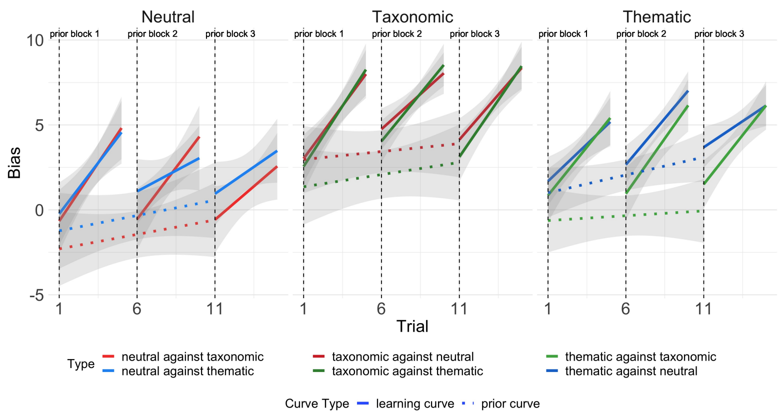


Figure 2. Learning and prior curves for each condition. Dashed vertical lines indicate the beginning of new experimental sets with new words. Learning curves reflect change of bias within each set. Prior curves reflect change of prior bias across three sets.

Higher education level was associated with a stronger thematic bias against both taxonomic and neutral choices. A closer investigation revealed that although thematic biases (against neutral and taxonomic choices) increased more with education, even at the highest education level, they did not surpass taxonomic biases. Taxonomic biases remained high across education levels. One possible reason is that taxonomic biases are easier to learn, and people become better at learning from diverse contexts to exhibit a broader ranger of biases (including thematic biases) with more education.

The only significant effect of age was in the neutral condition: older participants acquired less neutral bias against thematic choice versus against taxonomic choice (i.e., interactive effects between age and type), which provided tentative supporting evidence of the theory that elderly people re-increase their thematic bias (Smiley & Brown, 1979).

Table 1. Coefficient estimates for significant predictors. (*: $p < 0.05$,**: $p < 0.01$,***: $p < 0.005$, reference level of type: neutral against taxonomic)

Model	Variables	Predictor coefficient	95% CI
conditional	(Intercept)	1.98**	[0,78, 3.18]
learning	sequence order	1.59***	[1.18, 2.00]
effect	education	-2.00**	[-3.27, -0.73]
	type(taxonomic against neutral)	4.04***	[2.35, 5.73]
	type(taxonomic against thematic)	3.81***	[2.12, 5.50]
	type(thematic against neutral)	2.51**	[0.82, 4.19]
	education : type(thematic against neutral)	2.96**	[1.22, 4.69]
	education : type(thematic against taxonomic)	3.87***	[2.13, 5.60]
	age : type(neutral against thematic)	-1.02**	[-1.67, -0.38]
prior	set order	0.58**	[0.15, 1.02]
generalization	type(taxonomic against neutral)	4.70***	[2.54, 6.86]
effect	type(taxonomic against thematic)	3.34**	[1.18, 5.50]
	type(thematic against neutral)	3.45**	[1.29, 5.60]

In the prior generalization model, since there were no significant interactive effects, we removed these from the model and found a significant generalization effect: people increased their prior bias for the condition they were assigned to learn, and generalized the bias they learned from previous sets to a new set with novel lexical items (i.e., set order effects). We also found the same comparison type effects as found in the conditional learning effect model (Table 1).

4. Discussion

Our findings provide evidence supporting the hypothesis that exposure to different linguistic patterns cause individual differences in similarity biases. The observed learning effects across all conditions (e.g., taxonomic bias increases with more exposure to taxonomic co-occurrences) underscore that individuals indeed mold their similarity biases based on the word co-occurrence patterns they encounter. Furthermore, biases learned from language inputs can be generalized to novel lexical items, which suggests that exposure to language patterns in different co-occurrence contexts is a possible mechanism through which individual differences in similarity biases emerge. Altogether, our findings endorse the idea that our judgment of lexical and conceptual similarities can be guided by linguistic statistics, corroborating with prior studies that demonstrated such relationships in cross-cultural linguistic patterns and human similarity judgments (Le et al., 2023). Future research should investigate how different languages might naturally evolve to favor certain word co-occurrence contexts over others, to further shed light on how cross-cultural differences in similarity biases develop.

In conclusion, our study elucidates how linguistic patterns shape cognitive biases. The results underscore the importance of exploring language statistics exposure as a contributor to lexical representation and effective communication.

References

Borghi, A. M., & Caramelli, N. (2003). Situation bounded conceptual organization in children: From action to spatial relations. *Cognitive Development*, *18*(1), 49–60.

Brooks, P. J., Seiger-Gardner, L., & Sailor, K. (2014). Contrasting effects of associates and coordinates in children with and without language impairment: A picture–word interference study. *Applied Psycholinguistics*, *35*(3), 515–545.

Duan, Y., & Lupyan, G. (2023). Divergence in word meanings and its consequence for communication. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 45).

Hutchins, E., & Hazlehurst, B. (2006). How to invent a lexicon: the development of shared symbols in interaction. In *Artificial societies* (pp. 144–171). Routledge.

Ji, L., Zhang, Z., & Nisbett, R. E. (2004). Is it culture or is it language? examination of language effects in cross-cultural research on categorization. *Journal of Personality and Social Psychology*, *87*(1), 57-65.

Le, K. N., Gao, S., Frank, M. C., & Carstensen, A. (2023). Re-examining cross-cultural similarity judgments using language statistics. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 45).

Lin, E. L., & Murphy, G. L. (2001). Thematic relations in adults' concepts. *Journal of experimental psychology: General*, *130*(1), 3.

Luriia, A. R. (1976). *Cognitive development: Its cultural and social foundations.* Harvard university press.

Markman, E. M., & Hutchinson, J. E. (1984). Children's sensitivity to constraints on word meaning: Taxonomic versus thematic relations. *Cognitive Psychology*, *16*(1), 1-27.

Martí, L., Wu, S., Piantadosi, S. T., & Kidd, C. (2023). Latent diversity in human concepts. *Open Mind*, *7*, 79–92.

Nation, K., & Snowling, M. J. (1999). Developmental differences in sensitivity to semantic relations among good and poor comprehenders: Evidence from semantic priming. *Cognition*, *70*(1), B1–B13.

Nisbett, R. E., & Masuda, T. (2003). Culture and point of view. *Proceedings of the National Academy of Sciences*, *100*(19), 11163–11170.

Paridon, J. van, Liu, Q., & Lupyan, G. (2021). How do blind people know that blue is cold? distributional semantics encode color-adjective associations. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43).

Smiley, S. S., & Brown, A. L. (1979). Conceptual preference for thematic or taxonomic relations: A nonmonotonic age trend from preschool to old age. *Journal of Experimental Child Psychology*, *28*(2), 249–257.

Whitmore, J. M., Shore, W. J., & Smith, P. H. (2004). Partial knowledge of word meanings: Thematic and taxonomic representations. *Journal of Psycholinguistic Research*, *33*, 137–164.

Metaphor, emotion, and the evolved sensory interface

Frank H. Durgin*[1] and Annabella Boardman 2[1,2]

*Corresponding Author: fdurgin1@swarthmore.edu
[1]Department of Psychology, Swarthmore College, Swarthmore, PA, USA
[2]Department of Psychiatry, Brigham and Women's Hospital, Boston, MA, USA

Languages use figurative meanings derived from sensory experience. Sensory experience itself is an evolved interface rather than a true reporting of objective reality, and it contains a shared source of categories (e.g., warm, bitter, dark) upon which to ground language learning. Most sensory categories are additionally used to describe aspects of our social experience with usages that are similar across languages. Here we consider whether the content of these figurative usages is exclusively emotional. We used 99 concepts commonly metaphorized by 54 sensory/spatial adjectives. We measured the emotional content of those concepts using the semantic differential, and then used dual categorization tasks (IATs) to measure human conceptual alignments between concepts (N = 3405). Emotional content strongly predicted alignments, but significant additional alignment was found when two concepts shared a common sensory metaphor. Sensory metaphor conveys conceptual social information that goes well beyond emotional content.

1. Introduction

The human language faculty is a complex biological adaptation that evolved by natural selection (Pinker, 2003; Bloom & Pinker, 1990). There are large families of metaphorized lexical items in many languages that have both sensory meanings and social meanings. The use of sensory metaphor to describe persons is highly conserved across diverse languages (Asch, 1955, 1958). There has been a great deal of speculation concerning the foundational role of metaphor (and abstraction) in the evolution of human language (e.g., Cusky & Sommer, forthcoming; Ellison & Reinöhl, 2022; Smith & Höfler, 2014). Recent work on sensory metaphor has emphasized its use for conveying emotion (e.g., Citron & Goldberg, 2014), but emotional expression through non-verbal information is also highly conserved across many classes of terrestrial vertebrates (e.g., Congdon et al., 2019; Filippi et al., 2017; Lingle & Riede, 2014). The present study tested whether sensory metaphors leverage additional proto-conceptual information in the evolved sensory interface of humans (Hoffman, 2018; see also Zhu et al., forthcoming).

Here we show that concepts metaphorized by the same sensory metaphors show alignment strength that suggest unique conceptual contributions from shared evolved subjective categories, in addition to emotional content.

1.1. The Sensory Interface

Categories of subjective experience, including those for taste, smell, touch, sight and hearing, generally represent a shared source of mutual sensory understanding despite being inherently subjective. Even children can come to realize that there is no obvious way of knowing whether the appearance of “red” is the same from person to person, yet simply assuming a shared experience seems to work for language learning. For humans with the most common form of anomalous trichromatic color vision, for example, the consistency of difference in where color boundaries labels fall (e.g., green traffic lights appear pale green to those with deuteranomaly) provides additional confirmation that color experiences normally reflect an evolved 3-dimensional interface representing the ratios of activity across three cone types in consistent categories. Based on models of evolutionary processes, Hoffman (2018) has proposed that all perceptual categories are best construed as a kind of interface that summarily captures aspects of the world’s structure sufficient for survival and reproduction in ways that provide fitness without needing to be true, complete, or accurate.

1.2. Metaphorizing the Sensory Interface

The social psychologist, Solomon Asch, famously showed that certain kinds of personality descriptors colored other descriptors (1946). In particular, the words *cold* or *warm*, quite strongly changed the interpretation of the word *intelligent*. Less well known, Asch asked whether personality descriptors like warm and cold, which are sensory metaphors, vary from language to language (1955, 1958). He employed the aid of 6 experts in 6 languages from diverse language groups to collect evidence of the use of sensory metaphor in those languages in the descriptions of persons. Although usages varied in their details from language to language (e.g., a sour person might be one who had suffered a personal loss), Asch concluded that the similarities of use across languages were too prevalent to be accidental. Thus, Asch argued that seemingly unrelated languages chose similar sensory metaphors for recognizable social experiences.

A reductive interpretation of Asch’s finding might be that the emotional properties of sensory experiences were the only information that was being conveyed. For example, the basic tastes (salt, sweet, bitter, sour, savory) could be construed as varying in valence, arousal and dominance, and it might be these emotional properties that are primarily communicated by their use. Metaphors appear to convey emotion more strongly than literal counterparts (Citron & Goldberg, 2014). How might one differentiate sense-specific meaning from emotional communication?

1.3. Research Strategy

Here we report the result of an investigation in which we used psychological tools well-designed to implicitly measure conceptual alignment (the implicit association test or IAT, Greenwald et al., 1998), in combination with psychological tools designed to implicitly measure affective content (the Semantic Differential or *SemD*; Osgood et al., 1957). Although we expected that emotional content would explain a great deal of the variance when testing for alignment among concepts using the IAT (Xiong et al., 2006), we expected to find more specific effects of metaphoric alignment as well. That is, we expected that concepts that were commonly metaphorized by the same word (e.g., "smart" and "hurtful" can both be described as "sharp") might show conceptual alignments on the IAT above and beyond their measured emotional similarity.

A pre-registered pilot study using 7 sensory metaphors (each with 2 distinct meanings) measured conceptual alignments of 21 different concept pairs using IATs. Seven IATs paired concept pairs from same-metaphor sources and 14 used random pairings across metaphors. SemD ratings of the concepts were used to establish a 3-dimensional SemD score of emotional alignment. This pilot established that IAT scores could be predicted based on the correlation between the 3-dimensional SemD difference scores for each pair of words, and estimated that the small (non-significant) effect size of shared metaphoricity would require a much larger sample of items to reliably detect.

2. Methods

The present study was conducted online during the summer of 2022. Although not pre-registered, the analysis plan was developed as a replication of the pilot study which had pre-registered the exclusion criteria. Moreover, only a single planned (maximal) analysis was conducted. A total of 2783 (121*23) participants completed IATs while 522 (18*29) participants provided SemD ratings. About 18% of IAT participants were excluded for inattention (based on pre-registered criteria) and about 5% of ratings participants were excluded for poor attention, based on very low correlations between their ratings and the mean ratings for other participants.

2.1. Stimuli

A total of 99 concepts were selected based on distinct metaphoric senses of sensory or spatial adjectives. Different senses were operationalized as having a separate tab on thesaurus.com for that meaning and the meaning being figurative. For example, BRIGHT had tabs indicating figurative senses of *intelligent*, *promising*, and *cheerful*, among others. Thirty-three of the concepts were the sole metaphoric use listed for a sensory category, the other 66 concepts were groups of 2-6 figurative meanings from 21 sensory/spatial categories.

For each concept, an antonym appropriate for that metaphoric meaning was chosen by the authors from those provided by thesaurus.com in consultation with each other (e.g., *unintelligent*, *unpromising*, and *doleful*). Lists of synonyms (four for the target concept, and four for the antonym) were selected for each of the target pairs for use in IATs where the target concepts would be the category labels. The complete stimuli are available in supplemental online materials.

2.2. IATs

Concepts (and their antonyms) were paired as category labels either with concepts metaphorized by a *shared metaphoric source* (55 unique pairings were tested) or by a *different metaphoric source* (66 randomly-selected pairings). The latter pairings included 33 pairings among the concepts derived from sensory words that had at least two figurative meaning (*baseline*), and 33 pairings that crossed these with concepts from the list of sensory words with only a single figurative meaning (*control*). Thus, 121 unique IATs were created. Each consisted of the standard 7 blocks of trials, with 16 trials of practice at the two concepts separately, then 16 and 24 trials with the two mixed together, then 24 practice trials to reverse the labels on one of the categories, and then 16 and 24 trials with the two mixed together with the opposite alignment. The side of response, the order of alignment, and the dimension that switched were all randomized. On average, 19 participants were successfully tested with each of the 121 IATs. A D-score (difference in mean RTs divided by the pooled standard deviation) was computed for each participant. The experiment was run on PsyToolkit (Stoet, 2010, 2017); data were collected on Mechanical Turk via Cloud Research – i.e., TurkPrime (Litman et al., 2017).

2.3. Semantic Differential Data

Semantic differential ratings across 12 scales were collected for each target concept and its antonym. The concepts were divided across 18 different surveys to divide the labor. Participants had to complete a brief attention test prior to doing the ratings to ensure they were attentive. Ratings were averaged by concept and subjected to PCA (Dunteman, 1989) with normalized variables (scaled, centered) and orthogonal rotation. The first 3 components were consistent with *Evaluative*, *Potency*, and *Activity* dimensions normally identified by SemD procedures (corresponding to emotional valence, dominance, and arousal). Each concept/antonym pair was then assigned a 3D vector representing the difference between the two words along those three dimensions. The correlation between the vectors for concepts paired in our IATs was used as the *emotional alignment* predictor of IAT-assessed alignment (D-scores).

3. Results

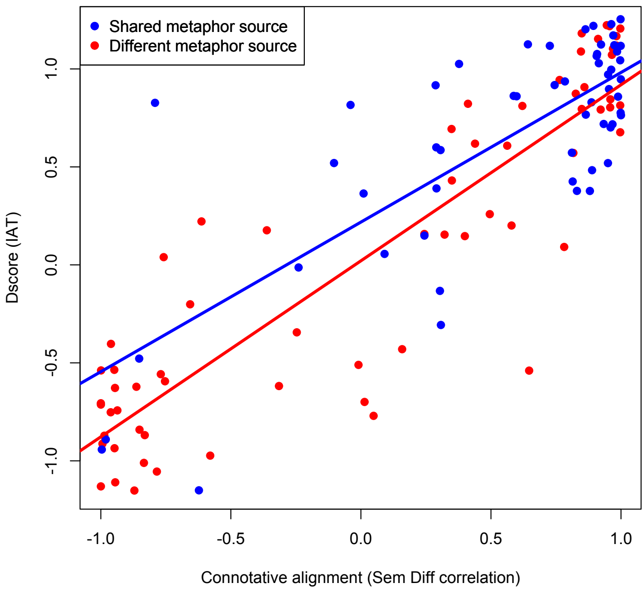


Figure 1. Mean IAT score (conceptual alignment) is plotted as a function of measured emotional alignment of concepts. Even with emotion taken into account, IAT scores are higher for concept pairs that are metaphorized with the same sensory word. Best fitting regression lines are shown.

Linear mixed effect regression (LMER) was used to analyze the D-scores for the IATs (computed using the advanced method Greenwald et al., 2003) with a maximal model, including both the *emotional alignment* predictor, and a three-level metaphor-source predictor representing the type of concept combination used (*baseline* random pairings vs. *shared metaphoric source* vs. *control* pairings). As expected, *emotional alignment* strongly predicted IAT scores, $\beta = 0.79$, $t(114.9) = 10.5$, $p < .0001$. However, there was also a significant effect of *shared metaphor source*, $\beta = 0.22$, $t(114.9) = 2.21$, $p = .029$. The results are shown by item in Figure 1, collapsing *control* and *baseline* pairings (which did not differ from each other) into a single category of *different metaphor source*.

4. Discussion

Abstraction, analogy, and the creation of figurative meaning are among the powerful drivers that have allowed human cognition to create categories that expand the classes of entities that words can refer to. The present results suggest that when people use sensory metaphor to convey more abstract meanings, they are tapping into information in our evolved perceptual interface that might be hard to otherwise articulate. The IAT appears to be particularly sensitive to emotional conceptual alignments among words, but also to more fine-grained meanings: Concepts metaphorized by the same sensory/spatial sources are, on average, more positively aligned than would be predicted based on their emotional content alone.

Recent criticisms of the IAT chiefly concern its use as a measure of individual differences (Schimmack, 2021). As a measure of cognitive patterns in populations, such as those shown here, it is like other group measures. When two categories are easily combined (align), switching the category alignment has a bigger cost, and this is what the IAT measures. Thus, in our data a high positive correlation in emotional content leads to a strongly positive IAT score, but a high negative correlation in emotional content leads to a strongly negative IAT score. Across all of this emotion content, however, there remains a strong positive shift in IAT measured alignment that suggests that sharing a metaphorized sensory meaning provides another form of category alignment. While it is possible that the semantic differential is simply inefficient at detecting emotional content, this seems unlikely to explain the current results.

The evolved sensory interface (Hoffman, 2018) that we experience as the directly-perceived world is a shared interface that provides a common sensory ground necessary for language learning. We can access abstract information in that shared sensory interface when we communicate about our social experiences using sensory words.

5. Supplementary Materials

Supplementary materials including all verbal stimuli used, and the full data set used for analysis, as well as the primary analysis code in R, are available online at https://osf.io/n47fk/?view_only=b00befa5bc074a5a8f0e09bc8a17b784.

Acknowledgements

This research was supported by a Swarthmore College Faculty Research Grant to FHD and a Frances Velay summer fellowship to AB.

References

Asch, S. E. (1946). Forming impressions of personality. *The Journal of Abnormal and Social Psychology, 41*(3), 258–290. https://doi.org/10.1037/h0055756

Asch, S. E. (1955). On the use of metaphor in the description of persons. In H. Werner (Ed.) *On Expressive Language* (pp. 29-38). Worcester: Clark University Press.

Asch, S. E. (1958). The metaphor: A psychological inquiry. In R. Tagiuri, & L. Petrullo (Eds.), *Person Perception and Interpersonal Behavior* (pp. 324-333). Stanford: Stanford University Press.

Citron, F. M., & Goldberg, A. E. (2014). Metaphorical sentences are more emotionally engaging than their literal counterparts. *Journal of Cognitive Neuroscience, 26*(11), 2585-2595. https://doi.org/10.1162/jocn_a_00654

Congdon, J. V., Hahn, A. H., Filippi, P., Campbell, K. A., Hoang, J., Scully, E. N., ... & Sturdy, C. B. (2019). Hear them roar: A comparison of black-capped chickadee (Poecile atricapillus) and human (Homo sapiens) perception of arousal in vocalizations across all classes of terrestrial vertebrates. *Journal of Comparative Psychology, 133*(4), 520. http://dx.doi.org/10.1037/com0000187

Cuskley, C., & Kees, S. (forthcoming). The evolution of linguistic iconicity and the cross-modal cognitive suite. To appear in O. Fisher, K. Akita, and P. Perniss (eds.), *Oxford Handbook of Iconicity in Language*. Oxford, UK: Oxford University Press.

Dunteman, G. H. (1989). *Principal components analysis* (No. 69). Newbury Park, CA: Sage.

Ellison, T.M., Reinöhl, U. (2022). Compositionality, Metaphor, and the Evolution of Language. *International Journal of Primatology*. https://doi.org/10.1007/s10764-022-00315-w

Filippi, P., Congdon, J. V., Hoang, J., Bowling, D. L., Reber, S. A., Pašukonis, A., ... & Güntürkün, O. (2017). Humans recognize emotional arousal in vocalizations across all classes of terrestrial vertebrates: Evidence for acoustic universals. *Proceedings of the Royal Society B: Biological Sciences, 284*(1859), https://doi.org/10.1098/rspb.2017.0990

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74(6)*, 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. Journal of personality and social psychology, 85(2), 197-216.

Hoffman, D. D. (2018). The interface theory of perception. *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience, 2*, 1-24. https://doi.org/10.1002/9781119170174.epcn216

Lingle, S., & Riede, T. (2014). Deer mothers are sensitive to infant distress vocalizations of diverse mammalian species. *The American Naturalist, 184(4)*, 510-522. https://doi.org/10.1086/677677

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods, 49*, 433-442. https://doi.org/10.3758/s13428-016-0727-z

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. University of Illinois Press.

Pinker, S. (2003). Language as an adaptation to the cognitive niche. *Studies in the Evolution of Language, 3*, 16-37.

Pinker, S., & Bloom, P. (1990). Natural language and natural selection. *Behavioral and Brain Sciences*, *13*, 707-784.

Schimmack, U. (2021). The Implicit Association Test: A method in search of a construct. *Perspectives on Psychological Science, 16*(2), 396-414. https://doi.org/10.1177/1745691619863798

Smith, A. D. M., & Höfler, S. H. (2014). The pivotal role of metaphor in the evolution of human language. In J. E. Diaz-Vera (Ed.) *Metaphor and Metonymy Across Time and Cultures: Perspectives on the Sociohistorical Linguistics of Figurative Language* (pp 123-139). Berlin, Germany: De Gruyter Mouton.

Stoet, G. (2010). PsyToolkit - A software package for programming psychological experiments using Linux. *Behavior Research Methods, 42*(4), 1096-1104. https://doi.org/10.3758/BRM.42.4.1096

Stoet, G. (2017). PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology, 44*(1), 24-31. https://doi.org/10.1177/0098628316677643

Xiong, M.J., Logan, G.D., & Franks, J.J. (2006). Testing the semantic differential as a model of task processes with the implicit association test. *Memory & Cognition, 34*, 1452–1463. https://doi.org/10.3758/BF03195910

Zhu, T. O., Chen, P., & Durgin, F. H. (forthcoming). The ups and downs of black and white: Do sensorimotor metaphors reflect an evolved perceptual interface? *Metaphor and Symbol.* http://dx.doi.org/10.1080/10926488.2024.2309977

What phonetics has to say about Neanderthal (*H. neanderthalensis*) speech capacities

Axel G. Ekström*[1] & Steven Moran[2,3]

*Corresponding Authors: axeleks@kth.se, steven.moran@unine.ch
[1]Division of Speech, Music & Hearing, KTH Royal Institute of Technology
[2]Institute of Biology, University of Neuchâtel
[3]Department of Anthropology, University of Miami

Neanderthals likely possessed some form of language, though it is uncertain whether their vocal anatomy allowed for the full range of modern human speech sounds. We synthesize literature on estimating Neanderthal speech capabilities and conclude that evidence supports the view that Neanderthals had restricted articulatory capacities compared to modern humans due to the shapes of their vocal tracts. To date, only two estimates of Neanderthal vocal tracts remain unrefuted – and both support the view that Neanderthals were limited with regards to the range of available speech sounds.

1. Introduction

In recent years, the prevailing view of Neanderthal linguistic abilities has shifted from a doltish species to beings with complex cognitive capacities (Dediu & Levinson, 2013). As summarized by Johansson (2015, p. 311), "from the consilience of evidence from anatomy, archeology, and DNA, one can conclude that some language abilities, if not necessarily full modern syntactic language, were present in Neanderthals." Here, we evaluate the contributions from phonetics-based approaches and estimates of Neanderthal speech capacities. The first widely discussed attempt to quantify Neanderthal phonetic capacities was performed by Lieberman and Crelin (1971), who suggested that Neanderthals were limited with regards to human-like speech production. Their findings are taken here as a baseline against which we evaluate more recent evidence arguing for, and against, the hypothesis that Neanderthals were "limited" from producing the full range of human speech sounds.

2. Basics of an evolutionary speech acoustics

In speech, the voice “source” from the vocal folds of the larynx is “filtered” in the supralaryngeal vocal tract (SVT) by the imposition of narrow constrictions using the various articulators, including the jaw, lips, velum, palate, and tongue (Fant, 1960). Because essential features of vocal anatomy are largely preserved across mammals, such fundamentals of speech acoustics have served as starting points for literature on the evolution of speech capacities (Negus, 1949; Lieberman et al., 1969, 1972; de Boer & Fitch, 2010; Fitch et al., 2016; Ekström & Edlund, 2023). The variable most crucial to the extent of the uniquely human range of speech sounds is the shape and position of the tongue inside the SVT, and the shape of the SVT itself (Lieberman et al., 1972; Carré et al., 1995; de Boer, 2010; de Boer & Fitch, 2010). In adult humans, the tongue root is descended into the pharynx, and the tongue, rounded in shape, is positioned in both the pharyngeal and oral cavities. The tongues of human infants and nonhuman mammals are flat in shape and contained almost wholly in the oral cavity (Negus, 1949). Resulting from disparate positions, while the principal musculature of the tongue is preserved across primates, innervation of equivalent musculature results in different vector forces in humans versus non-human primates (de Boer & Fitch, 2010). Resulting from a shortening of the face, and descent of the tongue root and concomitant descent of the larynx, the pharynx is markedly expanded in modern humans, resulting in roughly equal proportions between horizontal and vertical sections of the vocal tract (SVTh, SVTv). In chimpanzees, the SVTh is more than twice the length of the SVTv (Nishimura, 2005). Uniquely human proportions are optimal for generating a greatest-possible range of vowels (Carré et al., 1995; de Boer, 2010), allowing exploitation of the full range of speech sounds (Stevens, 1972). In particular, “point” vowels [a], [i], and [u] (the vowels in “ma”, “see”, and “do”) exhibit remarkable acoustic stability. For example, these vowels are uniquely recognizable even at high pitches (Friedrichs, 2017). Accordingly, the ability to articulate these speech sounds has received significant attention in relevant literature, with Lieberman and colleagues (1972) arguing that a uniquely human capacity to articulate these vowels reflected an evolutionary pressure for improved speech communication.

3. Estimates to date

3.1. The Negus–Keith estimates

To our knowledge, anatomist Victor Negus and anthropologist Arthur Keith were the first to attempt a reconstruction of Neanderthal supralaryngeal airways

(Negus, 1949). The authors concluded that the shapes of the Neanderthal tongue and pharynx would have been closer to those of the chimpanzee and human newborn than that of the adult modern human; this would imply support for the "limited Neanderthal" hypothesis. Unfortunately, these efforts are not sufficiently described in detail to allow for replication and will not be considered further here.

3.2. The Lieberman-Crelin estimates

The first modern estimates of Neanderthal speech capacities were performed by Lieberman and Crelin (1971) and Lieberman et al., (1972). These early efforts assumed that basicranial flexion provided a reliable indicator of the shape of vocal tracts. Specifically, Lieberman and colleagues argued that non-human primates, following from possessing short and narrow pharynges and flat-shaped tongues contained in the oral cavity, were effectively incapable of imposing the degrees of stricture necessary to achieve vowels [a], [i] and [u], which are all characterized by abrupt 10:1 discontinuities at the SVT midpoint, where SVTh and SVTv meet (Lieberman et al., 1972). Boë et al. (2002, p. 465–66) have incorrectly claimed the conclusions of the Lieberman/Crelin efforts were that Neanderthals "could not speak" and that an "increase in pharynx size [was a] necessary evolutionary preadaptation for speech". Rather, the Lieberman and Crelin estimates suggested that Neanderthal phonetic capacities, limited by a short and narrow pharynx, were less extensive than those of modern humans, with a vowel space that did not include the full extent of modern human vowels – but did include vowels [ɪ], [æ], and [ɛ] (the vowels in "bit", "cat" and "bed"). Results supported the view that Neanderthals may have been unable to articulate the full range of human speech sounds have since been a focal point in subsequent debate on Neanderthal speech capacities.

3.3. The Crelin estimates

Crelin (1987) extended the efforts begun by Lieberman and Crelin (1971) to various extinct hominids (see review in Ekström & Edlund, 2023). Also based on the "basicranial" assumption, Crelin determined that skulls of both australopiths and *H. habilis* were "apelike", while those of *H. erectus* were intermediate in form. Finally, based on a reconstruction of the "Steinheim skull" (an archaic human estimated to ~250–350 kya) that the species' vocal tract had been identical to that of a present-day *Homo sapiens* skull. The exact implications are somewhat ambiguous, as the taxonomic designation of the Steinheim skull individual has been subject to disagreement (see Stringer, 2016). It is, however, now generally considered an early Neanderthal lineage hominin. Accordingly, these later Crelin estimates would provide counterevidence against the early Lieberman and Crelin attempts, and against the "limited Neanderthal" hypothesis. However, more recent developments seemingly invalidate these earlier efforts.

3.4. A twist in the tale: The changing role of the basicranium

Efforts of Lieberman et al. (1972) and Crelin (1987) assumed that flexion of the skull base provided clues to the shape of species' vocal tracts. Human infants are born with "monkey vocal tracts" and basicranial angles, and achieve uniquely human proportions only later in life, once tongue root and larynx are sufficiently descended. Evidence emerging in the late 1990's showed that the tongue root and larynx of developing humans continue to descend even *after* cranial flexure has stabilized (Lieberman & McCarthy, 1999; Fitch & Giedd, 1999). These developments, thus, invalidate the assumption upon which earlier estimates were based. The marked flexion of the skull base does not provide the information necessary for determining the shapes of vocal tracts. This finding rendered the contribution of estimates based on this assumption ambiguous.

3.5. The Boë series

The Boë estimates are the only phonetics-based work to conclude that Neanderthals were "not morphologically handicapped for speech" (Boë et al., 1999, 2002). Several methodological constraints make this determination problematic, however. The Boë series are the only estimates to base their work on the "basicranial flexure" assumption that were published after the publication of results that invalidate it. Paradoxically, Boë et al. cite Lieberman and McCarthy (1999), who invalidate the assumptions upon which their work is based. More significantly, however, the algorithm employed by the authors preserves the tongue shapes of the modern humans upon which those shapes were based – in the words of de Boer and Fitch (2010, p. 42), "precisely the aspect of the anatomy that is in question" (see also Lieberman, 2007, 2012). The same method would show that any animal would possess the full range of human speech: accordingly, the Boë series cannot be taken as evidence that Neanderthals were not "handicapped for speech".

3.6. The Barney estimates

Barney et al. (2012) provide a novel estimation method and attempt to qualify speech capacities in fossil specimens. The authors present a case study based on the (relatively recent) "La Ferrassie" skull (dated to ~50kya), and report a range of possible values, including displacement of both jaw and hyoid from anatomically predicted locations; however, neither resultant vowel space extends to that of their modern human referent. While the authors do not present a systematic exploration of results of the method as applied to other Neanderthal specimens, the study provides support for the “limited Neanderthal” hypothesis.

3.7. The McCarthy series

The most exhaustive series of estimates to date were performed by McCarthy (Lieberman & McCarthy, 2007; Lieberman, 2007, 2012) (Table 1). Namely, it is in theory possible to estimate a position for the hyolaryngeal complex, necessary for achieving a “roughly equal” SVTh-SVTv relationship (presumed necessary for the full extent of human vowel space) at resting state conditions. The McCarthy estimates indicated that, in order to achieve roughly 1:1 SVTh–SVTv proportions, the larynx of Neanderthals, reflecting a combination of short necks and long faces, would have to be placed inside the thorax – an “impossible” configuration that is not found in any extant primate: “the short neck and long Neanderthal SVTh would place the cricoid cartilage behind the sternum, permitting human speech but precluding eating” (Lieberman, 2007, p. 47). Neanderthals would, accordingly, be unable to produce “fully modern” speech.

Table 1. Summary of results of estimates.

Effort	*Neanderthals limited?*	*Refuted?*	*Source of refutation*
Negus reconstruction	*YES*	N/A	*Insufficiently described*
Lieberman/Crelin estimates	*YES*	*YES*	*McCarthy and Lieberman (1999); Fitch and Giedd (1999)*
Crelin series	*NO*	*YES*	*McCarthy and Lieberman (1999); Fitch and Giedd (1999)*
Boë series	*NO*	*YES*	*de Boer and Fitch (2010)*
Barney estimate	*YES*	*NO*	N/A
McCarthy series	*YES*	*NO*	N/A

4. What the hyoid cannot tell us about speech

On various occasions, the shape of Neanderthal hyoid bones has been claimed to be indicative of speech capacities. For example, Frayer (2017, p. 236) claims a hyoid enables “a full appreciation of the modern language capacities of Neanderthals.” This argument is, however, inconsistent with the science of speech production, and does not recognize that the crucial variable for phonetic capacities is the shape of the SVT. Quoting Lieberman (1999, p. 175): “An isolated Neanderthal hyoid bone can’t tell you whether the Neandertal had a human vocal tract, because the hyoid bone and larynx descend as children mature, without any systematic change in shape.” Hyoid shape alone, thus, does not inform researchers of phonetic range available to extinct hominids. Any relationship between hyoid and phonetic capacities relies on soft tissue reconstruction (McCarthy & Lieberman, 2007; Barney et al., 2012). For the claim, “We now know that… the Neanderthal vocal tract is capable of producing vowels very similar or identical to modern Europeans”, Frayer (2017, p. 235) cites the Barney estimates (which indicate the opposite), and work by Dediu and Levinson (2013) who base their arguments to this effect on the refuted Boë estimates.

5. What hearing cannot tell us about speech

Conde-Valverde et al. (2021, p. 609) argue, based on reconstructions of Neanderthal auditory anatomy and the assumption that “the occupied bandwidth [computed based sound power transmission] is directly related to the efficiency of the vocal communication”, that “Neanderthals and Homo sapiens had similar auditory and speech capacities.” The authors do not, however, provide any evidence directly bearing on vocal anatomy. In addition, novel evidence suggests that auditory thresholds emerged prior to the human-chimpanzee split (Stoessel et al., 2023). The contribution toward supporting or refuting the “limited Neanderthal” hypothesis is thus uncertain.

6. Conclusions

We have synthesized decades of work informed by acoustic phonetics bearing on Neanderthal speech capacities. To date, only one estimate (Boë et al., 1999) has concluded that Neanderthals were “not morphologically handicapped for speech” – and this work has been firmly refuted (de Boer & Fitch, 2010). Other evidence purported to indicate speech capacities – the shape of Neanderthal hyoids and inferred auditory capacities – are not useful for this purpose. The history of hominin vocal tract estimates is clouded with novel findings invalidating earlier work, and future efforts may reveal as-yet unknown relationships bearing on vocal tract shapes of extinct hominids. Currently, however, available speech acoustics research supports the view that, while Neanderthals likely possessed language, they may have been limited to a less extensive range of speech sounds.

Acknowledgements

We gratefully acknowledge funding from the Swiss National Science Foundation (PCEFP1_186841; SM). The results of this work and the tools used will be made more widely accessible through the national infrastructure Språkbanken Tal under funding from the Swedish Research Council (2017-00626).

References

Barney, A., Martelli, S., Serrurier, A., & Steele, J. (2012). Articulatory capacity of Neanderthals, a very recent and human-like fossil hominin. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1585), 88-102.

Boë, L. J., Maeda, S., & Heim, J. L. (1999). Neandertal man was not morphologically handicapped for speech. *Evolution of Communication*, *3*(1), 49-77.

Boë, L. J., Heim, J. L., Honda, K., & Maeda, S. (2002). The potential Neandertal vowel space was as large as that of modern humans. *Journal of Phonetics, 30,* 465-484.

Carré, R., Lindblom, B., & MacNeilage, P.F. (1995). Rôle de l'acoustique dans l'évolution du conduit vocal humain. *Comptes Rendus de l'Académie des Sciences*, Série IIb, 320, 471–476.

Conde-Valverde, M., Martínez, I., Quam, R. M., Rosa, M., Velez, A. D., Lorenzo, C., ... & Arsuaga, J. L. (2021). Neanderthals and Homo sapiens had similar auditory and speech capacities. *Nature ecology & evolution*, *5*(5), 609-615.

Crelin, E. S. (1987). *The human vocal tract: Anatomy, function, development, and evolution*. Vantage Press.

de Boer, B. (2010). Modelling vocal anatomy's significant effect on speech. *Journal of Evolutionary Psychology*, *8*(4), 351–366.

de Boer, B., & Fitch, W. T. (2010). Computer models of vocal tract evolution: An overview and critique. *Adaptive Behavior*, *18*(1), 36–47.

Dediu, D., & Levinson, S. C. (2013). On the antiquity of language: the reinterpretation of Neandertal linguistic capacities and its consequences. *Frontiers in psychology*, *4*, 397.

Ekström, A. G., & Edlund, J. (2023). Evolution of the human tongue and emergence of speech biomechanics. *Frontiers in Psychology*, *14*(1150778). Doi: 10.3389/fpsyg.2023.1150778.

Ekström, A. G. (in press). Phonetic potential of primate vocal tracts and the legacy of Philip Lieberman (1934–2022). *American Journal of Primatology*.

Fant, G. (1960). *The acoustic theory of speech production*. Mouton: The Hague.

Fitch, W. T., & Giedd, J. (1999). Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *The Journal of the Acoustical Society of America*, *106*(3), 1511-1522.

Frayer, D. W. (2017). Talking hyoids and talking Neanderthals. *Human Paleontology and Prehistory: Contributions in Honor of Yoel Rak*, 233-237.

Friedrichs, D. (2017). *Beyond formants: vowel perception at high fundamental frequencies* (Doctoral dissertation, University of Zurich).

Johansson, S. (2015). Language abilities in Neanderthals. *Annu. Rev. Linguist., 1*(1), 311-332.

Lieberman, D. E., & McCarthy, R. C. (1999). The ontogeny of cranial base angulation in humans and chimpanzees and its implications for reconstructing pharyngeal dimensions. *Journal of human evolution*, *36*(5), 487-517.

Lieberman, D. E., McCarthy, R. C., Hiiemae, K. M., & Palmer, J. B. (2001). Ontogeny of postnatal hyoid and larynx descent in humans. *Archives of oral biology*, *46*(2), 117-128.

Lieberman, D. E. (2011). *The evolution of the human head*. Harvard University Press.

Lieberman, P., & Crelin, E. S. (1971). On the speech of Neanderthal man. *Linguistic inquiry*, *2*(2), 203-222.

Lieberman, P., Crelin, E. S., & Klatt, D. H. (1972). Phonetic ability and related anatomy of the newborn and adult human, Neanderthal man, and the chimpanzee. *American Anthropologist*, *74*(3), 287–307.

Lieberman, P. (2007). The evolution of human speech: Its anatomical and neural bases. *Current anthropology*, *48*(1), 39–66.

Lieberman, P., & McCarthy, R. (2007). Tracking the evolution of language and speech: Comparing vocal tracts to identify speech capabilities. *Expedition: The magazine of the University of Pennsylvania*, *49*(2), 15-20.

Lieberman, P. (2012). Vocal tract anatomy and the neural bases of talking. *Journal of Phonetics*, *40*(4), 608–622.

Negus, V. E. (1949). *Comparative anatomy and physiology of the larynx*. Heinemann.

Nishimura, T. (2005). Developmental changes in the shape of the supralaryngeal vocal tract in chimpanzees. *American Journal of Physical Anthropology*, *126*(2), 193-204.

Stevens, K. N. (1972) The quantal nature of speech: Evidence from articulatory-acoustic data. In (Eds.), E. E. David, Jr & P. B. Denes, *Human communication: a unified view* (pp. 51–66). McGrawHill.

Stoessel, A., David, R., Bornitz, M., Ossmann, S., & Neudert, M. (2023). Auditory thresholds compatible with optimal speech reception likely evolved before the human-chimpanzee split. *Scientific Reports*, *13*(1), 20732.

Stringer, C. (2016). The origin and evolution of Homo sapiens. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1698), 20150237.

Phonetic properties of chimpanzee, gorilla, and orangutan hoots tell a uniform story and point to new frontiers

Axel G. Ekström[1*], Lara Nellissen[2,3], Tatiana Bortolato[4,5,6], Catherine Crockford[4,5,6], Jens Edlund[1], Shelly Masi[3], Klaus Zuberbühler[2], Adriano R. Lameira[7], Roman W. Wittig[4,5,6], Sven Grawunder[4,8], and Steven Moran[2,9]

[*]Corresponding Author: axeleks@kth.se
[1]Speech, Music & Hearing, KTH Royal Institute of Technology, Stockholm, Sweden
[2]Institute of Biology, University of Neuchâtel, Neuchâtel, Switzerland
[3]Ecoanthropologie (EA), Centre National de la Recherche Scientifique/Museum National d'Histoire Naturelle, University Paris Diderot, Sorbonne Paris Cite, Musee de l'Homme, Paris, France.
[4]Department of Human Behavior, Ecology and Culture, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
[5]Ape Social Mind Lab, Institut des Sciences Cognitives, Lyon, France
[6]Taï Chimpanzee Project, Centre Suisse de Recherches Scientifiques, Ivory Coast
[7]Department of Psychology, University of Warwick, Coventry, UK
[8]Department of Speech Science and Phonetics, University of Halle–Wittenberg, Halle, Germany
[9]Department of Anthropology, University of Miami, Coral Gables, FL, US

We present a first–ever comparison of phonetic properties across vocalizations by great apes. We show that "hoot–like" calls by (males of) all non–human great ape genera – chimpanzees, gorillas, and orangutans – overlap with those of human back rounded vowels. Our work underlines the importance of studying the production of calls. Observations from both comparative vocal morphology (non–human great apes have short–and–narrow pharynges and tongues contained in the oral cavity) and observations of vocalizing animals indicate they likely achieve these qualities with disparate articulatory gestures.

1. Introduction

All nonhuman great apes produce hoot–like calls, but intra-species comparisons of call properties are rare in the literature. Here, we present the first three–way comparison of properties of hoot-like calls produced by all extant nonhuman great ape genera: chimpanzees (*Pan* .spp), gorillas (*Gorilla* .spp), and orangutans (*Pongo* .spp). We note that across species, hoot-like calls exhibit comparable and overlapping properties; we further suggest that this apparent uniformity results from a derived feature in great apes, i.e., the employment of protruding and rounded lips in call production. We apply the terminology of phonetics and refer to apparent spectral peaks as formants.

2. Methods

2.1. *Formant estimation*

To estimate formants, we applied the PREQUEL protocol (Ekström, Moran, Sundberg, & Lameira, 2023). Fundamental frequency (f_0) was assessed visually by hand using correlograms (Granqvist & Hammarberg, 2003), and apparent first formant–second formant (F_1, F_2) coordinates were synthesized and matched for f_0 and compared to the original recording.

2.2. *Samples*

2.2.1. *Chimpanzees*

Pant hoot calls by Western chimpanzees (*Pan troglodytes verus*) (N=50, 5 individuals) were recorded by TB at the Taï Chimpanzee Project, Ivory Coast. Pant hoots are divided into four phases, with breathy early-bout lower-frequency vocalizations gradually transitioning into open-mouth high-f_0 screams (climaxes) (Grawunder et al., 2022). Because higher-frequency calls are generally nonconducive to formant analysis (Ekström, 2023), only introduction and build-up phases were examined. Pant hoots are performed on both inhalation and exhalation (inbreath, outbreath); we limited analysis to utterances on exhalations (Eklund, 2008). For this study, all sampled individuals were males. This was to control for possible effects of sexual dimorphism – although in comparison with gorillas and orangutans, chimpanzees exhibit relatively little dimorphism (Dixson, 1998).

2.2.2. *Gorillas*

Western gorilla (*G. gorilla*) silverbacks hoots (N=34, 2 individuals). Data were recorded by LN at the Bai Hokou and Mongambe field sites in the Dzanga-Sangha Protected Areas in the Central African Republic. Hoots analyzed here are from two adult silverbacks each leading an independent group.

2.2.3. *Orangutans*

Bornean flanged male orangutan (*P. pygmaeus wurmbii*) long calls (Lameira & Wich, 2008) were sampled and analyzed (N=109, 9 individuals). Calls were collected at the Tuanan Orangutan Research Station, Central Kalimantan, Borneo, Indonesia by ARL. In our sample data, because recording quality was variable with higher frequencies being lost to high-frequency noise (e.g., birdsong), it was often necessary to segment calls where select portions showed clear and consistent formant frequencies (Ekström et al., 2023).

3. Results

Results of our investigations (Figure 1, Table 1) show there is substantial overlap between phonetic properties of great ape hoot–like calls (chimpanzee pant hoot "hoo's", gorilla soft hoots, orangutan long calls) between all three species, and with human close back rounded vowel [u].

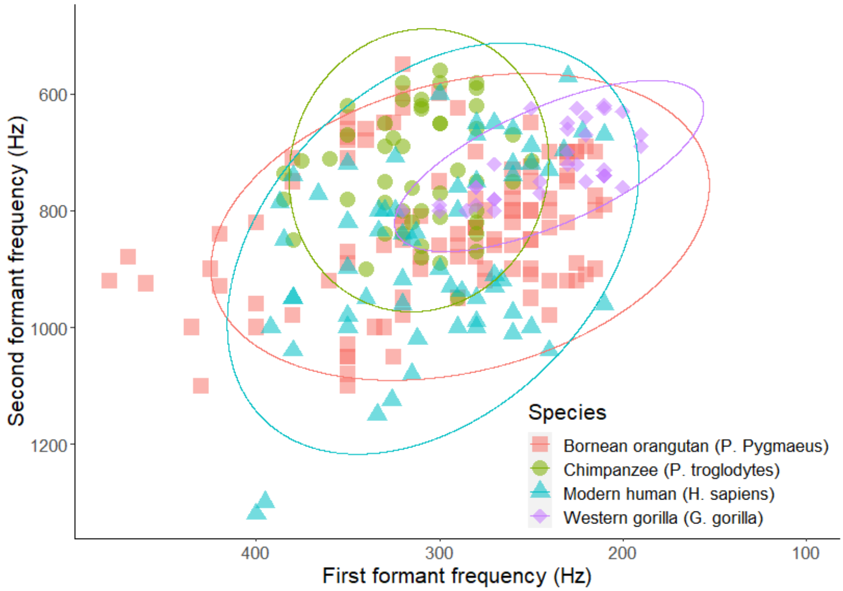


Figure 1. Vowel-like spaces of chimpanzees, gorillas, orangutans producing hoots, and modern human males speaking [u] (Peterson & Barney, 1952). All calls produced by adult males.

Table 1. Mean estimated formants for chimpanzee and gorilla hoots, orangutan long calls, and male human [u]. Values in Hertz. Standard deviations in parentheses.

	Chimpanzee hoots	Gorilla hoots and soft hoots	Orangutan long calls	Human [u]
F_1	334 (73)	240 (35)	299 (63)	305 (50)
F_2	748 (108)	721 (60)	829 (120)	871 (159)

We ran linear mixed model analyses in R (ver. 4.3.1) using the *lme4* package (Bates, Mächler, Bolker, & Walker, 2015). We included response as the dependent variable and added random effects of individual animal. Significance was calculated using the *lmerTest* package (Kuznetsova, Brockhoff, & Christensen, 2017), which applies Satterthwaite's method to estimate degrees of freedom and generate *P*-values for mixed models. The model specifications were: $F_n \sim$ Species + (1 | Subject). Data were log10 transformed prior to analysis. For both F_1 and F_2, analyses yielded highly significant intercepts (P<.001). For F_2, the effect of the chimpanzee data was statistically significant (P=.012); there were no other significant factor settings. The analysis findings are summarized in Table 2 and Table 3.

Table 2. Linear mixed model for F_1. Reference level is human [u].

	Estimate	Standard error	df	t	P
(Intercept)	306.12	8.67	55.79	35.32	.00***
Chimpanzee	9	21.46	35.75	.42	.68
Gorilla	–61.41	30.96	29.82	–1.98	.06
Orangutan	15.70	16.76	35.68	.94	.36

Table 3. Linear mixed model for F_2. Reference level is human [u].

	Estimate	Standard error	df	t	P
(Intercept)	875.91	24.45	47.42	35.82	.00***
Chimpanzee	–167.84	63.84	37.96	–2.62	.01*
Gorilla	–138.91	94.43	34.41	–1.47	.15
Orangutan	2.66	49.91	37.71	.05	.96

4. Discussion

4.1. *Disparate vocal tract lengths*

Our results suggest that there is substantial overlap between phonetic properties of hoot-like calls across great apes, while species' calls may still be categorically distinguishable from each other. To verify commonalities, we would optimally seek to scale our estimated formants according to vocal tracts length for each species. However, for orangutans and gorillas, no reliable estimates exist. We may, however, take note of some reported findings. Goldstein (1980) measured the length of pharyngeal cavity in an adult human male at 8.9 cm, and length of the oral cavity at 8.1 cm, for a total vocal tract length (VTL) of approximately 17 cm. Nishimura (2005) estimated a total vocal tract length for an adult male chimpanzee at 18.12 cm (computed by adding lengths reported for vertical and horizontal portions of the tract): in terms of total length, the longer chimpanzee face and oral tract compensates for the short pharyngeal tract. Vocal tract lengths of adult humans and chimpanzees thus largely overlap. Further, adult male chimpanzees and orangutans overlap in body size (Dixson, 1998). Because vocal tract length is intimately correlated with body size across primates (Fitch, 1997) – including great apes (Nishimura, 2005) – we may tentatively assume that vocal tract lengths of chimpanzee and orangutan males are likely to overlap also. For these species, biases resulting from differences in vocal tract length are likely to be relatively minor. Notably, however, gorillas represent a significant exception to this trend, with upper height boundaries of an upright silverback measuring some 30 cm above that of an adult male chimpanzee (Dixson, 1998). Thus, it is likely that while vocal tract lengths for adult male chimpanzees and orangutans may overlap, this is markedly less likely so for silverbacks. Provisioning of great ape vocal tract length data – in particular for gorillas – would significantly improve upon the possibilities to draw conclusions from our findings.

4.2. *Air sacs*

Another imposition to vocal tract length scaling are laryngeal air sacs, the acoustic consequences of which are contested in the literature. de Boer (2012) has suggested air sacs shift up resonances under 2kHz, and introduce an additional low-frequency resonance. We argue that further acoustic modeling efforts, particularly those aimed at exploring the interaction between air sacs and protruding rounded lips, may help resolve this incongruity. Visual inspection of vocalizing animals would also facilitate the modeling of these articulatory behaviors. If air sacs indeed introduce an additional low-frequency resonance, we should treat the apparent F_2 as a "shifted-up" F_1, and F_1 as a novel resonance induced by the presence of the sacs. This may be consistent with our data. Namely, assuming a VTL = 18.12 cm (Nishimura, 2005), predicted F_1–F_2 of schwa are F_1 = 487 Hz, F_2 = 1461 Hz, according to :

$$Fn = (2n - 1) \cdot c/4 \cdot L \tag{1}$$

where *n* is the n_{th} formant, *c* is the speed of sound, and *L* is the total length of the tract. Assuming that the articulatory gestures observed by Parr et al. (2005) and Grawunder et al. (2022) for hoots are accurate, we would assume a *longer* tract, as the lips are protruded, effectively shifting down all formants. For example, at VTL = 22 cm, we would expect F_1 = 401 Hz, and F_2 = 1203 Hz, assuming a uniform tube. This is, however, definitively inconsistent with our observations, which put F_2 some $\approx$450 Hz below this estimate. Consistent with with studies of human vowel production, many studies – including recent efforts by Grawunder et al. (2022) and Ekström et al. (2023) – have focused on measuring and reporting F_1–F_2 dispersion. However, the categorization scheme reported in Grawunder et al. (2022) indicate that, as in human speakers, the apparent first spectral peak is tied to jaw height. Moving forward, we argue that it is necessary to report at least the first three apparent formants, so as to definitively either support or refute the purported roles of air sacs.

4.3. *Articulation*

From an evolutionary perspective, our findings are intriguing, as related vocal tract anatomy and morphology differs significantly between human and nonhuman primates, with nonhuman primates possessing narrow oro– and laryngopharynges

(Negus, 1949, p. 196); the homologous structure in humans is elongated, and open in [u]. There are also likely limitations on intraoral gestures resulting from tongue morphology (Takemoto, 2008). Thus, nonhuman great ape production of [u]-like calls likely involves disparate articulatory gestures (de Boer & Fitch, 2010), likely affecting speech potential more broadly (Ekström & Edlund, 2023a). In addition, speech acoustics modeling exercises indicate that a “two–tube” vocal tract (with proportionate pharyngeal and oral tracts) is more efficient than the standard primate vocal tract (Carré, Lindblom, & MacNeilage, 1995). Observations of apes producing hoots indicate that such calls are often (though not always) produced with comparatively extreme contortions of the lips (Parr, Cohen, & Waal, 2005). Elongating an acoustic chamber will always shift down formants (Fant, 1960), and indeed in chimpanzees, such gestures have been shown to be associated with a reliable shifting down of formants (Grawunder et al., 2022). Our data suggest that the articulatory gestures employed in hoot production – that is, the tendency to affect formant dispersions via the elongation and/or narrowing of the lip passage through rounding – may be a derived feature in nonhuman primates. Understanding morphological aspects involved in the production of apparently similar vowel qualities and vowel-like qualities may yield important insights into critical developments ultimately facilitating the evolution of speech in human ancestors (Ekström & Edlund, 2023b).

Acknowledgements

The results of this work and the tools used will be made more widely accessible through the national infrastructure Språkbanken Tal under funding from the Swedish Research Council (2017–00626). SM and AE were funded through the Swiss National Science Foundation (PCEFP1_186841) and AL by the UK Research & Innovation, Future Leaders Fellowship (grant agreement number MR/T04229X/1; ARL). We also thank the Ministère de l’Enseignement supérieur et de la Recherche scientifique; des Eaux et Forêts en Côte d’Ivoire and the Office Ivoirien des Parcs et Réserves, and the Ministry of Higher Education and Scientific Research of Central African Republic (CAR) for permission to conduct this research, the Dzanga–Sangha Protected Areas and WWF CAR for allowing us to carry out fieldwork at their sites, and the UMR 7206 of the National Museum of Natural History in Paris.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Carré, R., Lindblom, B., & MacNeilage, P. F. (1995). Rôle de l’acoustique dans l’évolution du conduit vocal humain. *Comptes Rendus de l’Académie des Sciences, Série IIb*, *320*, 471–476.

de Boer, B. (2012). Loss of air sacs improved hominin speech abilities. *Journal of Human Evolution*, *62*, 1–6.

de Boer, B., & Fitch, T. W. (2010). Computer models of vocal tract evolution: An overview and critique. *Adaptive Behavior*, *18*, 36–47.

Dixson, A. F. (1998). *Primate sexuality: Comparative studies of the prosimians, monkeys, apes, and human beings.* Oxford, UK: Oxford University Press.

Eklund, R. (2008). Pulmonic ingressive phonation: Diachronic and synchronic characteristics, distribution and function in animal and human sound production and in human speech. *Journal of the International Phonetic Association*, *38*, 235—324.

Ekström, A. G. (2023). Ape vowel-like sounds remain elusive: A comment on Grawunder et al. (2022). *International Journal of Primatology*, *44*, 237—239.

Ekström, A. G., & Edlund, J. (2023a). Evolution of the human tongue and emergence of speech biomechanics. *Frontiers in Psychology*, *14*, 2132.

Ekström, A. G., & Edlund, J. (2023b). Sketches of chimpanzee (*Pan troglodytes*) hoo's: vowels by any other name? *Primates*, *65*, 81–88.

Ekström, A. G., Moran, S., Sundberg, J., & Lameira, A. R. (2023). Prequel: Supervised phonetic approaches to analyses of great ape quasi-vowels. In R. Skarnitzl & J. Volín (Eds.), *Proceedings of the 20th international congress of phonetic sciences* (pp. 3076–3080). Prague, Czech Republic: GUARANT International spol. s r.o.

Fant, G. (1960). *The acoustic theory of speech production.* The Hague: Mouton.

Fitch, T. W. (1997). Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. *The Journal of the Acoustical Society of America*, *102*, 1213–1222.

Goldstein, U. G. (1980). *An articulatory model for the vocal tracts of growing children.* [Doctoral Dissertation]. Massachusetts Institute of Technology.

Granqvist, S., & Hammarberg, B. (2003). The correlogram: A visual display of periodicity. *The Journal of the Acoustical Society of America*, *114*, 2934–2945.

Grawunder, S., Uomini, N., Samuni, L., Bortolato, T., Girard-Buttoz, C., R. M. Wittig, R. M., & Crockford, C. (2022). Chimpanzee vowel-like sounds and voice quality suggest formant space expansion through the hominoid lineage. *Philosophical Transactions of the Royal Society B*, *377*, 20200455.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26.

Lameira, A. R., & Wich, S. A. (2008). Orangutan long call degradation and individuality over distance: A playback approach. *International Journal of Primatology*, *29*, 615–625.

Negus, V. E. (1949). *Comparative anatomy and physiology of the larynx.* Heinemann.

Nishimura, T. (2005). Developmental changes in the shape of the supralaryngeal vocal tract in chimpanzees. *American Journal of Physical Anthropology*, *126*, 193–204.

Parr, L. A., Cohen, M., & Waal, F. D. (2005). Influence of social context on the use of blended and graded facial displays in chimpanzees. *International Journal of Primatology*, *26*, 73–103.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, *24*, 175–184.

Takemoto, H. (2008). Morphological analyses and 3d modeling of the tongue musculature of the chimpanzee (pan troglodytes). *American Journal of Primatology*, *70*, 966–975.

The phoneme as cognitive technology

Caleb Everett[1], Steven Moran[1,2], and Axel G. Ekström[3*]

[*]Corresponding Author: axeleks@kth.se
[1]Department of Anthropology, University of Miami, Coral Gables, FL, US
[2]Institute of Biology, University of Neuchâtel, Neuchâtel, Switzerland
[3]Division of Speech, Music & Hearing, KTH Royal Institute of Technology, Stockholm, Sweden

1. Introduction

While the number and characteristics of vowels and consonants are highly variable across the world's spoken languages (Moran & McCloy, 2019), all speakers make consistent and deliberate use of a relatively narrow set of contrastive speech sounds, i.e., phonemes. Such remarkable ubiquity is suggestive of extensive benefits to their speakers. Importantly, however, the phoneme is not a naturally occurring phenomenon. Rather, we argue that phonological systems constitute cognitive tools, i.e., that they support, guide, and extend speaker cognitive capacities (Everett, 2017). We make several claims toward this point.

2. Information rate

Compared with the communication systems of nonhuman animals, the consistent and socially deliberated use and reuse of phonemes enables rapid information transmission rates through syllabic speech (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). Based on their work constructing reading machines for the blind, Liberman and colleagues noted that if spoken language were produced letter by letter (or phoneme by phoneme), then speech rates would be significantly reduced. In real-life speech, however, speech is always coarticulated – the production of a phoneme is continually affected by its context. Phonemes in a spoken language may serve as points of reference, even as speech sounds are distorted by linguistic and extra-linguistic (e.g., emotional speech) contexts (Lindblom, 1990). The cultural "invention" of the phoneme in human evolution and society, thus, enabled the rapid information transmission rates universally observed across human languages (Coupé, Oh, Dediu, & Pellegrino, 2019).

3. The phoneme as developmental scaffold

Native-language input provides developing human infants with acoustic-perceptual goals (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1996), the replication or quasi-replication of which, serve as a marker of emergent social consciousness and identity (Tomasello, 2003). This development takes place in all developing human infants within the first few years of life (Vihman, 2014), likely exploiting existant subcortical neuromotor systems (see review in Ekström, 2022). Experimental evidence from perceptual abilities by human infants suggests that already by six months of age, humans begin selectively discriminating between language-specific phonemes; this selective perception eventually develops into a so-called "perceptual magnet" (Kuhl, 1991), effectively serving as a prototype for its category.

4. Perceptual overlap

Non-human animal oral tracts afford the capacity to non-uniformly affect formants. Recent work indicates that orangutan "long calls" are readily perceived as phonemic by listeners (Ekström, Moran, Sundberg, & Lameira, 2023). We propose that exposure to systems of speech sounds bias human perceptual systems toward selective perception of environmental sounds as speech-like, including the calls of other animals. Accordingly, vocalizations produced even by distantly related animals such as domestic cats (*Felis catus*) are uniformly transcribed – across even unrelated languages – as a consonant-vowel-consonant or consonant-vowel-vowel sequence corresponding, e.g., to /miauw/. We argue that the reason animal vocalizations may be perceived as essentially "word–like" (Nicastro & Owren, 2003) is contingent on phonemic learning.

5. Implications for evolutionary phonology

Treating contrasting sounds of speech as products of culture may also open up novel discussions of its evolution. Recent work in archaeology and anthropology point to distinctions between findings, such that tools of relatively low complexity cannot be used to infer cultural transmission (Snyder, Reeves, & Tennie, 2022). We suggest that similar distinctions may be made with regards to systems of speech in human evolution, such that a possible "early" system could be independently invented by individuals and groups, while more complex systems required cultural transmission (Benítez-Burraco & Kempe, 2018). This suggestion may be explored through computational modeling (Kirby & Hurford, 2002). Our view complements cognitive linguistic perspectives on human perception and consciousness by emphasizing the sounds of speech themselves.

Acknowledgements

SM and AE were funded by the Swiss National Science Foundation (PCEFP1_186841). The results of this work and the tools used will be made more widely accessible through the national infrastructure Språkbanken Tal under funding from the The Swedish Research Council (2017-00626, AE).

References

Benítez-Burraco, A., & Kempe, V. (2018). The emergence of modern languages: has human self-domestication optimized language transmission? *Frontiers in Psychology*, *9*, 551.

Coupé, C., Oh, Y. M., Dediu, D., & Pellegrino, F. (2019). Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science Advances*, *5*(9), eaaw2594.

Ekström, A. G. (2022). Motor constellation theory: A model of infants' phonological development. *Frontiers in Psychology*, *13*, 996894.

Ekström, A. G., Moran, S., Sundberg, J., & Lameira, A. R. (2023). Prequel: Supervised phonetic approaches to analyses of great ape quasi-vowels. In R. Skarnitzl & J. Volín (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences* (pp. 3076–3080). Prague, Czech Republic.

Everett, C. (2017). *Numbers and the making of us: Counting and the course of human cultures.* Harvard University Press.

Kirby, S., & Hurford, J. R. (2002). The emergence of linguistic structure: An overview of the iterated learning model. In A. Cangelosi & D. Parisi (Eds.), *Simulating the evolution of language* (pp. 121–147). Springer London.

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, *50(2)*, 93–107.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1996). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, *255(5044)*, 606–608.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431–461.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the h & h theory. In W. J. Hardcastle & A. Marchal (Eds.), (pp. 403–439). Springer.

Moran, S., & McCloy, D. (2019). *Phoible 2.0.* Jena: Max Planck Institute for the Science of Human History.

Nicastro, N., & Owren, M. J. (2003). Classification of domestic cat (*Felis catus*) vocalizations by naive and experienced human listeners. *Journal of Comparative Psychology*, *117(1)*, 44.

Snyder, W. D., Reeves, J. S., & Tennie, C. (2022). Early knapping techniques do not necessitate cultural transmission. *Science Advances*, *8(27)*, eabo2894.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition.* Cambridge, MA: Harvard University Press.

Vihman, M. M. (2014). *Phonological development: The first two years.* Wiley.

Locating the emergence of hominin pragmatic competence

Andrew Feeney
andrew.feeney@northumbria.ac.uk
Department of Humanities, Northumbria University, Newcastle upon Tyne, NE1 8ST, United Kingdom

Interest in language has a recorded history stretching back to the ancient Greeks, and, as Chomsky rightly notes, 'the traditional description of language as sound with a meaning [is] traceable at least back to Aristotle' (1995: 2). In the intervening years, the study of 'grammar' encompassed inquiry into linguistic sounds and their patterns (phonetics and phonology), the internal structure of meaningful units (morphology), the relationships between these units (syntax) and the encoding of meanings (semantics). P*ragmatics*, the study of how speakers rely on shared contextual information in communication, was a lamentably late addition to the field of linguistic inquiry and the domain was only afforded a name in the last century (Morrison, 1938) and for much of its subsequent existence was considered an adjunct to the mainstream: a 'wastebasket' for problematic phenomena in Bar-Hillel's (1971) terminology.

In recent years, pragmatic competence has begun to be, rightly, seen as an indispensable component in understanding the evolution of language. However, a number of questions arise before the nature of the pragmatic role in language evolution can be fully established. The most fundamental of these concerns the relative temporal location of the emergence of pragmatic competence, and its consequences, in the interpretation and generation of linguistic structure.

It is now generally agreed that linguistic code is semantically underspecified and that, at the very least, effective communication relies upon pragmatic enrichment (Grice, 1967; Sperber & Wilson, 1986). Currently though, there is still disagreement among even those who place the role of pragmatics at the centre of the human capacity for language, as to the nature of its role, and thus evolutionary history. On the one hand, there is a school of thought which maintains that a 'proto presumption of relevance' emerged in precursor species prior to that of any form of language (e.g. Scott-Phillips, 2014; Scott-Phillips & Heintz, 2023). An alternative position posits the existence of an intermediate stage of pragmatic competence beyond the synthesis of simple immediate contextual information, but lacking the cognitive complexity of processing implicature (e.g. Bar-On, 2021). Finally, some (e.g. Carston, in prep) stress the primacy of hierarchical, structured syntax as a necessary impetus to kick start the pragmatic process required for the production and comprehension of modern language.

In all cases, pragmatic competence relies upon a number of pre-requisites including, firstly, a substantial willingness to cooperate in the exchange of information, and, following from this, a cognitive bias among interlocuters towards the presumption of shared relevance. A very rudimentary form of the first of these appears attested in the behaviour of chimpanzees and bonobos who (alone among primates) appear to have some capacity for latent collaboration (Melis et al., 2006; Gibson, 2012). Furthermore, they appear able to apply this basic cooperation to communication, and are capable of using context to determine the meaning of ambiguous gestures (Graham, in prep). However, evidence for more substantial cooperation only begins to be seen around 1.9 million years ago (mya) in the hominin clade with the appearance of *Homo erectus* (Tomasello, 2008; Tomasello et al., 2012) and species' specific concomitant cultural developments including the production and use of the first mode 2 tools (Beyene et al., 2013), coordinated hunting and scavenging, the first exodus out of Africa and possibly the controlled use of fire for the processing of food (Wynn, 2012; Wrangham, 2009). As this period is also associated with the earliest evolutionary adaptations for vocalisation, it is not implausible to posit the advent of protolanguage during the subsequent 500 thousand years during which these innovations arose (Bickerton, 2009; Tallerman, 2012).

However, the question that then surfaces is why, if these early hominins were cooperative and tuned to the presumption of relevance (and even had protolanguage), there was almost complete cultural stasis in the million years or so that followed from around 1.5 mya. As the archaeologist J. Desmond Clark is reported to have observed, if *H. erectus* had (proto)language then 'these ancient people were saying the same thing to each other, over and over and over again' (Stringer, 2011: 125). Whatever evolutionary adaptations had been bestowed upon this species, possibly including a vocal protolanguage, they were incapable of initiating a second punctuation of hominin equilibrium, which occurs only around 500 thousand years ago (kya). Pragmatic competence, in the form of 'expression unleashed' (Heintz & Scott-Phillips, 2023) alone, while necessary, was not sufficient for the transition to modern language. The problem is resolved if we conclude that at this point hominins underwent another major neuro-cognitive development that resulted in an enhanced cognition, again attested by the relatively sudden appearance of cultural changes. This cognitive machinery ultimately gives rise to multi-order intentionality, and only at this stage did the extant pragmatic (and other linguistic) capacities become utilised in the development of modern language.

References

Bar-Hillel, Y. (1971). Out of the Pragmatic Wastebasket. *Linguistic Inquiry,* 2 (3), 401-407.

Bar-On, D. (2021) How to do things with nonwords: pragmatics, biosemantics, and origins of language in animal communication. *Biology & Philosophy, 36* (6), 50.

Beyene, Y., Katoh, S., WoldeGabriel, G., Hart, W. K., Uto, K., Sudo, M., et al. (2013) The characteristics and chronology of the earliest. Acheulean at Konso, Ethiopia. *Proceedings of the National Academy of Sciences of the United States of America, 110*(5), 1584-1591.

Bickerton, D. (2009) *Adam's tongue : how humans made language, how language made humans*. New York: Hill and Wang.

Carston, R. (in prep) Pragmatics, Proto-Ostension and the Language-Communication Relation

Chomsky, Noam. (1995) *The Minimalist Program*. Cambridge, MA.: MIT Press

Gibson, K., Rita. (2012) Language or protolanguage? A review of the ape language literature. In M. Tallerman & K. R. Gibson (Eds.), *The Oxford handbook of language evolution* Oxford: Oxford University Press.

Graham, K. E. (in prep) Pragmatic Primates

Graham, K. E., Furuichi, T., & Byrne, R. W. (2020). Context, not sequence order, affects the meaning of bonobo (Pan paniscus) gestures. *Gesture, 19* (2-3), 335-364.

Grice, H. P. (1967). Logic and Conversation. In: Grice (1989) *Studies in the Ways of words*. Cambridge, MA: Harvard University Press.

Heintz, C., & Scott-Phillips, T. (2023). Expression unleashed: The evolutionary and cognitive foundations of human communication. *Behavioral and Brain Sciences, 46,*

Melis, A. P., Hare, B., & Tomasello, M. (2006) Engineering cooperation in chimpanzees: tolerance constraints on cooperation. *Animal Behaviour, 72,* 275-286.

Morrison, C. W. (1938). Foundations of the Theory of Sign. In O. Neurath, R. Carnap & C. Morris (Eds.), *International Encyclopaedia of Unified Science* 77-138. Chicago: University of Chicago Press.

Scott-Phillips, T. (2014). *Speaking our minds: why human communication is different, and how language evolved to make it special*. Palgrave Macmillan.

Scott-Phillips, T., & Heintz, C. (2023) Animal Communication in Linguistic and Cognitive Perspective. *Annual Review of Linguistics, 9* (1), 93-111.

Sperber, D., & Wilson, D. (1986). *Relevance: communication and cognition*. Cambridge, Mass.: Harvard University Press.

Stringer, C. (2011) *The origin of our species*. London: Allen Lane.

Tallerman, M. (2012) Protolanguage. In: M. Tallerman & K. R. Gibson (Eds.), *The Oxford handbook of language evolution* Oxford: Oxford University Press.
Tomasello, M. (2008) *Origins of human communication*. Cambridge, Mass.: MIT Press.
Tomasello, M., Melis, A. P., Tennie, C., Wyman, E., & Herrmann, E. (2012) Two Key Steps in the Evolution of Human Cooperation The Interdependence Hypothesis. *Current Anthropology, 53*(6), 673-692.
Wrangham, R. W. (2009) *Catching fire : how cooking made us human,* New York: Basic Books.
Wynn, T. (2012) The Palaeolithic Record. *In:* Tallerman, M. & Gibson, K. R. (eds.) *The Oxford Handbook of Language Evolution.* Oxford: Oxford University Press.

Exploring the influence of word definition on AI-generated facial representations

Vojtěch Fiala[*1], Przemyslaw Zywiczynski[1], Anna Szala[1], Marek Placiński[1], Slawomir Wacewicz[1]

[*]Corresponding Author: vfiala@umk.pl
[1]Center for Language Evolution Studies, Nicolaus Copernicus University, Toruń, Poland

One of the key questions in language evolution is the role of linguistic labels in driving cognitive evolution, esp. by influencing categorisation, including visual perceptual stimuli (Lupyan 2006, Lupyan & Casasanto 2015). Interestingly, this problem finds parallels in a large swathe of human-evolutionary research since a standard study format in evolutionary psychology is to rate visual stimuli – very often human faces – on characteristics presented as verbal labels (Langlois et al., 2000). Even closely related psychological concepts are defined by different words and cued by distinctive facial configurations (e.g., Mileva, 2016). For example, a face imagined as "socially attractive" (a face appealing in the social context) may possess different characteristics than a "sexually attractive face" (Kruger, 2006). Our study therefore explores the problem of the interface between facial characteristics and their linguistic descriptions.

Despite the impact that a linguistic formulation may exert on perception (Lupyan et al. 2020), studies on perceived facial characteristics, e.g., attractiveness, usually do not provide a definition of the focus characteristic they want the participants to rate: out of 65 on the topic from 2021–2023, we found only 1 paper that did this. While a majority of these articles concern mate choice, others focus on social psychology, economics, and political sciences. Given this broad diversity of contexts in which 'attractiveness' was studied, we decided to test whether a deep learning text-to-image model, Stable Diffusion XL, returns detectably different images when prompted for sexual vs social attractiveness. Trained on ~6 billion image-text pairs, the model presents a potent source of images accompanied by name descriptions. We used one prompt for sexual attractiveness (*prompt 1*: "attractive European man/woman, sexual mating and partnership context") and another for social attractiveness (*prompt 2*: "attractive prosocial, friendly, and cooperative European man/woman" with "sexy" and "good mate" as negative prompts).

For each of the two prompts, we generated 120 facial images (60 women, 60 men), and applied standard selection criteria in evolutionary psychology: full face visible, not horizontally or vertically tilted, closed mouth, neutral expression, and facial features not covered by hair/facial hair, resulting in 108 faces (54 W, 54 M). The faces were subject to geometric morphometrics analysis. We marked each face with 72 landmarks (see Kleisner et al., 2019) in the program tpsDig (Rohlf, 2015). We ran a generalised Procrustes analysis (separately for each sex) using package geomorph in R (Baken et al. 2021; Adams, 2023), and computed the average configuration for each category (sexual vs. social, Figure 1). We then calculated the distance of each facial configuration from its ingroup/outgroup mean configuration. A subsequent analysis with the function permudist of the package Morpho (Schlager, 2017) suggests that the faces tend to cluster around its groups' average configuration (Procrustes distance between social-sexual group means [PDM] = 0.00030, $p < 0.001$ for men, PDM = 0.00032, $p < 0.001$ for women). Although numerically small, the difference between groups is significant.

The results suggest that AI-generated representations of human faces (for this preliminary study limited to the faces of European origin) systematically differ depending on the particular phrasing of the prompt, to the degree that AI-generated socially vs. sexually attractive faces can be distinguished by their geometric-morphometrics properties. This novel approach to visualising the variance of human facial characteristics based on their linguistic description can be flexibly applied to the faces of people from other ethnic backgrounds and can be extended with the application of AI graphic tools (currently, we are using the text-to-image generator Dall-e, and faces of models from populations outside Europe). Our future research aims to include automated estimates of facial attractiveness, real human faces, human raters, and different facial characteristics to test whether both human-made and automated facial characterisations are sensitive to word definitions.

Figure 1. Representations of sexually (I) vs socially (II) attractive faces of 'men' (left) and 'women' (right), based on images created by Stable Diffusion.

Acknowledgements

We are thankful to prof. Karel Kleisner for introducing us to the methods of geometric morphometrics.

Short summary:
The role of linguistic labels in categorisation and determining the scope of a category may have large consequences for ascribing characteristics to visual stimuli. Despite that, studies on perceived facial characteristics, including attractiveness, usually do not provide the readers with a definition of the rated category. Using the deep learning text-to-image model Stable Difusion, we created average visual representations of two distinctive definitions of facial attractiveness: socially and sexually attractive faces. Subsequently, we used tools of landmark-based geometric morphometrics to explore if the visual representations differ across the definitions. The results suggest that the AI-generated representations of attractive human faces are context- and verbal description-dependent and that faces generated with social and sexual attractiveness primes can be distinguished by methods of geometric morphometrics. This result is relevant in the context of links between the perceptual salience of facial characteristics and linguistic expressions referring to these characteristics.

OSF link to the project: https://osf.io/qund8/?view_only=c5654a1ffe5c497e840702b00cf123ec

References

Adams, D., Collyer, M., Kaliontzopoulou, A., & Baken, E. (2023). Geomorph: Software for geometric morphometric analyses. R package version 4.0.6. https://cran.r-project.org/package=geomorph.

Baken, E. K., Collyer, M. L., Kaliontzopoulou, A., & Adams, D. C. (2021). geomorph v4. 0 and gmShiny: Enhanced analytics and a new graphical interface for a comprehensive morphometric experience. *Methods in Ecology and Evolution*, *12*(12), 2355–2363.

Kleisner, K., Pokorný, Š., & Saribay, S. A. (2019). Toward a new approach to cross-cultural distinctiveness and typicality of human faces: The cross-group typicality/distinctiveness metric. *Frontiers in psychology*, *10*, 124.

Kruger, D. J. (2006). Male facial masculinity influences attributions of personality and reproductive strategy. *Personal Relationships*, *13*(4), 451–463.

Langlois, J. H., Kalakanis, L., Rubenstein, A. J., Larson, A., Hallam, M., & Smoot, M. (2000). Maxims or myths of beauty? A meta-analytic and theoretical review. *Psychological bulletin, 126*(3), 390.

Lupyan, G. 2006. Labels Facilitate Learning of Novel Categories. In: Cangelosi, A., Smith, A.D.M, & Smith, A. (eds.) 2006. The Evolution of Language. Proceedings of the 6th International Conference (EVOLANG6). Singapore: World Scientific Publishing.

Lupyan, G., & Casasanto, D. (2015). Meaningless words promote meaningful categorization. *Language and Cognition*, 7(2), 167-193.

Lupyan, G., Rahman, R. A., Boroditsky, L., & Clark, A. (2020). Effects of language on visual perception. *Trends in cognitive sciences*, 24(11), 930-944.

Mileva, V. (2016). Social status in humans: Differentiating the Cues to dominance and prestige in men and women. Dissertation Thesis. University of Stirling

Rohlf, F. J. (2015). The tps series of software. *Hystrix, 26*(1), 9–12.

Schlager, S. (2017). Morpho and Rvcg–Shape Analysis in R: R-Packages for geometric morphometrics, shape analysis and surface manipulations. In Statistical shape and deformation analysis (pp. 217–256). *Academic Press*.

What's the deal with large language models? A Comparative Evolutionary Perspective

Molly Flaherty*[1], Christine Cuskley[2]

*Corresponding Author: moflaherty@davidson.edu
[1]Department of Psychology, Davidson College, Davidson NC, USA
[2]Language Evolution, Acquisition, and Development Group, Newcastle University, Newcastle upon Tyne, UK

Text generated by large language models (LLMs) is now often indistinguishable from text generated by humans. Designers of these models claim they are perilously close to reaching human-like levels of general intelligence (e.g. Bengio et al., 2023), with progress rapidly advancing as models become "multimodal" (though note that this is confined to the ability to integrate text and images, and does not approach the extent of multimodality in human language; Goldin-Meadow, 1999; Rasenberg et al, 2022). Alongside this, some cognitive scientists have declared that either impressive LLM performance (Contreras Kallens et al., 2023; Piantadosi, 2023; Frank, 2023) or its specific shortcomings (Katzir, 2023; Chomsky et al., 2023) provides compelling new evidence for longstanding debates about domain specificity and innateness of language (see Pleyer & Hartmann, 2018, for a summary of these debates in language evolution particularly).

The current work begins by questioning the immediate relevance of LLMs for understanding human language. We situate the language capacity of LLMs in a comparative perspective with the human language capacity using a Tinbergian framework of mechanisms, development, adaptive function and phylogeny. While LLMs share narrow adaptive function with human language (with coverage for producing text only, which is a proxy of only some spoken languages with written forms), the way in which an LLM develops its language capacity ("ontogeny") and the mechanisms by which it stores linguistic representations differ fundamentally. Despite the technically shared ancestry of human language and LLMs, we situate the phylogeny of LLMs as being a case of an *analogous* trait to human language, representing (if anything) partially overlapping convergent evolution, accomplished by means of human design rather than natural selection.

While this places the behaviour of LLMs as fundamentally different from human language in many important respects, it nonetheless provides us with one of the first close *functional* comparators for human language. The performance of LLMs alone cannot tell us much about how the human language faculty works, but careful probing of the difference in performance between humans and LLMs on specific, language evolution-oriented tasks provides an opportunity for new insights.

To this end, we report tests of a series of Artificial Language Learning (ALL) tasks focusing on training on partial systematic vocabularies with structured meaning spaces (adapted from Kirby, Cornish, Tamariz and Smith, 2014) followed by a test on full vocabulary. We contrast multimodal language models with text only models. Text only models were relatively successful in learning labels for seen items and recalling the meanings of novel words (unseen items). Some multi-modal models had comparable performance for learning labels, but when asked to generate descriptions of images for novel words results were inconsistent. A subset of multimodal models also showed uneven performance, often “collapsing” a systematic vocabulary by reproducing the same label repeatedly, even for unrelated shapes - an underspecification found in early iterated ALL without communication pressures (Kirby, Cornish & Smith, 2008). While the source of this disparity remains unclear, we suggest that the vision models generally used in multimodal LLMs are ill-adapted to dealing with the kinds of simple geometric images often deliberately chosen for ALL with humans. However, this same simplicity makes these images easy to describe systematically to text only LLMs.

While the reason why multimodal models fail to capture simple geometric structures in images may be obvious (it likely reflects the predominantly photographic input on which these multimodal models were trained), the implications of this are nonetheless interesting. Humans also have predominantly complex visual input, particularly during our evolutionary history, and yet simple geometric shapes are used in ALL studies with humans precisely because they make structure in meaning spaces readily apparent. We report on ongoing work using text to image generation models to create structured photographic image sets for systematically testing ALL across multimodal models and humans.

References

Bengio, Y., Russel, S., Musk, E., Wozniak, S., et al. (2023, March 22). Pause Giant AI Experiments: An Open Letter. Future of Life Institute. https://futureoflife.org/open-letter/pause-giant-ai-experiments/ Accessed 13 October, 2023.

Chomsky, N., Roberts, I., & Watumull, J. (2023, March 8). Opinion: The false promise of ChatGPT. *New York Times*.

Contreras Kallens, P., Kristensen-McLachlan, R. D., & Christiansen, M. H. (2023). Large Language Models Demonstrate the Potential of Statistical Learning in Language. *Cognitive Science*, *47*(3), e13256. https://doi.org/10.1111/cogs.13256

Frank, M. (2023). Bridging the data gap between children and large language models. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2023.08.007

Goldin-Meadow, S. (1999). The role of gesture in communication and thinking. *Trends in Cognitive Sciences*, *3*(11), 419-429.

Hurford, J. R. (2011). *The origins of grammar: Language in the light of evolution II*. OUP Oxford.

Katzir, R. (2023). Why large language models are poor theories of human linguistic cognition. A response to Piantadosi (2023). Lingbuzz preprint. https://lingbuzz.net/lingbuzz/007190

Piantadosi, S. (2023) Modern language models refute Chomsky's approach to language. Lingbuzz preprint. https://lingbuzz.net/lingbuzz/007180

Pleyer, M., & Hartmann, S. (2019). Constructing a Consensus on Language Evolution? Convergences and Differences Between Biolinguistic and Usage-Based Approaches. *Frontiers in Psychology*, *10*. https://www.frontiersin.org/articles/10.3389/fpsyg.2019.02537

Rasenberg, M., Pouw, W., Özyürek, A., & Dingemanse, M. (2022). The multimodal nature of communicative efficiency in social interaction. *Scientific Reports*, *12*(1), 19111.

Chain shifts and transphonologizations are driven by homophony avoidance

Stefon Flego[1], Andrew Wedel[*2]

[*]Corresponding Author: wedel@arizona.edu
[1]Department of English, Virginia Tech, Blacksburg, USA
[2]Department of Linguistics, University of Arizona, Tucson, USA

Loss of a phoneme contrast through merger is significantly associated with a low degree of resulting word-level homophony (Kaplan, 2011; Wedel et al., 2013). For example, there are very few English words distinguished by the two low-back vowels /ɑ/ as in 'cot', and /ɔ/ as in 'caught', and this vowel contrast has merged in many dialects of North American English. In contrast, phonemes that do not merge are characterized by many such 'minimal pairs', that is, words that would become homophonous if the phonemes were to merge. In these previous studies, homophony avoidance was associated with *lack* of change. Here, we show that homophony avoidance appears to also drive two superficially distinct, *active* sound changes. **Chain shifts** occur when a set of phonemes move in concert within phonetic space. For example, the front vowels in New Zealand English have undergone a chain shift upwards, such that the vowel /æ/ in ‘pat’ has raised to /ɛ/, and the original /ɛ/ in ‘pet’ has raised to /e/ (Bauer et al., 2007; Hay et al., 2015). **Transphonologizations**, on the other hand, occur when the primary cue distinguishing two phonemes merges, while a minor cue expands in concert to become the primary cue. For example, aspirated and lenis stops in Korean are historically distinguished by a voice-onset-time (VOT) difference, with a minor distinction in f0 on the following vowel. In modern Seoul Korean, this VOT difference is collapsing, while the f0 difference has expanded to become the primary cue (Silva, 2006). These two superficially distinct classes of sound change have in common that lexical contrast is maintained throughout the change: in a chain shift, one phoneme moves into the space occupied by another, which concomitantly shifts away into a neighboring part of the phonetic space. In a transphonologization, one cue to a phoneme contrast merges, while at the same time another cue to the same contrast expands. Here we show that while phoneme mergers are characterized by a low number of minimal pairs (and therefore low numbers of resulting homophones), chain shifts and transphonologizations are characterized by especially high numbers of minimal pairs. Our dataset comprises a genetically and areally diverse set of twelve languages which have undergone historically recent mergers, chain shifts and transphonologizations. We identified the number of minimal pairs distinguished by each phoneme contrast participating in a change, as well the number of minimal pairs associated with a comparison set of similar contrasts that have not participated in a change. Relative to the distribution of minimal pairs of non-changing phoneme contrasts, we find that mergers, as shown previously, are drawn significantly from the lower end of this distribution. Conversely, we find that contrasts that have undergone chain-shifts and transphonologizations are drawn significantly from the higher end of this distribution (Figure 1). These findings are consistent with computational work showing that category shift in one phonetic dimension (e.g., chain shifts) and category shift across multiple phonetic dimensions (e.g., transphonologizations), can be driven by the same mechanism (Wedel, 2012). More broadly, these findings provide support for usage-based theories of change in which information-theoretic factors like homophony avoidance play a fundamental role in shaping languages' sound systems over time (Flego, 2022; Sóskuthy, 2013; Winter & Wedel, 2016; Winters et al., 2018).

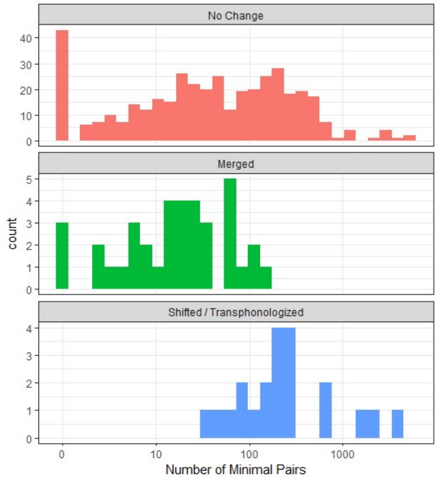


Figure 1. Comparison of minimal pair counts for each sound change category.

References

Bauer, L., Warren, P., Bardsley, D., Kennedy, M., & Major, G. (2007). New Zealand English. *Journal of the International Phonetic Association*, *37*(1), 97-102.

Flego, S. (2022). *The emergence of vowel quality mutation in Germanic and Dinka-Nuer: Modeling the role of information-theoretic factors using agent-based simulation* [Doctoral dissertation, Indiana University Bloomington].

Hay, J. B., Pierrehumbert, J. B., Walker, A. J., & LaShell, P. (2015). Tracking word frequency effects through 130 years of sound change. *Cognition, 139*, 83-91.

Kaplan, A. (2011). How much homophony is normal?. *Journal of Linguistics*, *47*(3), 631-671.

Silva, D. J. (2006). Acoustic evidence for the emergence of tonal contrast in contemporary Korean. *Phonology*, *23*(2), 287-308.

Sóskuthy, M. (2013). *Phonetic biases and systemic effects in the actuation of sound change* [Doctoral dissertation, University of Edinburgh].

Wedel, A. (2012). Lexical contrast maintenance and the organization of sublexical contrast systems. *Language and Cognition*, *4*(4), 319-355.

Wedel, A., Kaplan, A., & Jackson, S. (2013). High functional load inhibits phonological contrast loss: A corpus study. *Cognition, 128*(2), 179-186.

Winter, B., & Wedel, A. (2016). The co-evolution of speech and the lexicon: The interaction of functional pressures, redundancy, and category variation. *Topics in Cognitive Science, 8*(2), 503-513.

Winters, J., Kirby, S., & Smith, K. (2018). Contextual predictability shapes signal autonomy. *Cognition, 176*, 15-30.

Communicative efficiency and social biases modulate language learning in autistic and allistic individuals

Lauren Fletcher[*1], Hugh Rabagliati[1], and Jennifer Culbertson[1]

[*]Corresponding Author: l.e.f.fletcher@sms.ed.ac.uk
[1]University of Edinburgh, Edinburgh, UK

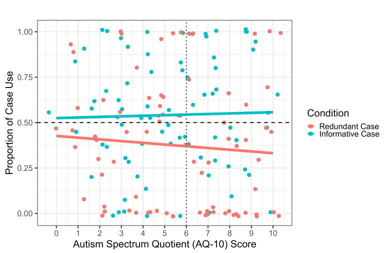
Communicative efficiency has been cited as driving many core features of linguistic systems, through a trade-off between communicative accuracy and production effort (Zipf, 1935; Roberts & Fedzechkina, 2018; Kurumada & Jaeger, 2015; Fedzechkina & Jaeger, 2020). This trade-off has been appealed to in explaining the relationship between whether a language uses case to mark grammatical roles and whether it has a fixed word order (Sinnemäki, 2008). Specifically, it has been argued that, where fixed word order alone is enough of a cue to grammatical role assignment, redundant marking with case is inefficient as it requires unnecessary production effort. On the other hand, where word order is flexible, the extra effort to produce case is warranted in order to maintain communicative accuracy. The role of social biases in modulating this trade-off have been investigated by Roberts and Fedzechkina (2018) and Fedzechkina, Hartley, and Roberts (2022), who find that learners are willing to put in more production effort, or sacrifice communicative accuracy, to meet social goals.

The study of communicative efficiency has, however, largely assumed an important degree of homogeneity amongst language users. Yet, it is increasingly clear that the assumption of homogeneity is incorrect, and that speakers vary in a multitude of ways, including in terms of neurotype. In the context of evolutionary linguistics, it is important to note that the majority of neurodivergent individuals are active members of their language communities and thus differences among individuals with different neurotypes could have an impact on language evolution as a whole. We focused on the impact of a specific neurotype – autism – on the relationship between social biases and communicative efficiency. We chose this because autism is formally characterised as a social-communicative developmental disorder (American Psychiatric Association, 2013), and differences in social and communicative skills are a key hallmark of day-to-day life for most autistic people. For example, many autistic people perform what is known as 'masking', where they hide their autistic traits in order to facilitate conversation with non-autistic people (Hull et al., 2017; Cook, Crane, & Mandy, 2023; Pearson & Rose, 2021)

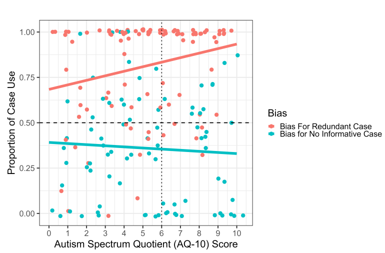
In Experiment 1, we sought to determine whether autistic people displayed the communicative efficiency trade-off at all with regards to the negative relationship between fixed word order and case marking. We replicated the ‘no-bias’ conditions of Roberts and Fedzechkina (2018) and Fedzechkina et al. (2022) in both the autistic and non-autistic populations. In these conditions, participants learnt a simple artificial language in which object case was marked 50% of the time. In the informative case condition, word order was flexible, with 50% use of SOV and 50% use of OSV. In the redundant case condition, word order was fixed, with 100% use of SOV. We found that autistic people re-structured their input to be communicatively efficient in the same way as their allistic peers. Autistic people reduce the use of case in the redundant case condition, whilst they retain the use of case in the informative case condition, to the same degree as allistic people.

In Experiment 2, we introduced social biases into our paradigm by partially replicating the ‘bias’ conditions of Roberts and Fedzechkina (2018) and Fedzechkina et al. (2022). In the *bias for redundant case* condition, participants were told to favour a group of aliens who used object case marking 100% of the time in a fixed word order language. In the *bias for no informative case* condition, participants were told to favour a group of aliens who *did not* use object case marking in a flexible word order language. In this case, we found a clear difference between participants based on neurotype: autistic people in the *bias for redundant case condition* were more likely to increase their use of redundant case, despite it costing effort to produce, in order to meet a social bias. We argue that this reflects the fact that autistic people may put more effort into social situations through strategies such as masking in order to compensate for the difficulties they face in social interactions with allistic people.

Our results underscore the importance of considering the impact of neurotype in language evolution. In this case, our results illustrate that the strength of the effect of social biases varies across the population in ways that may impact language change. More generally, though, these results indicate that neurotype may interact in significant ways with the kinds of cognitive biases and mechanisms we appeal to in language evolution research.



(a) Case use across the two no bias conditions and its interaction with AQ-10 score.



(b) Case use across the two bias conditions and its interaction with AQ-10 score

Acknowledgements

This work is in part supported by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1), by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 757643 to the second author), and by funding from the University of Edinburgh, School of Informatics and School of Philosophy, Psychology, and Language Sciences, and the Centre for Language Evolution at the University of Edinburgh.

References

American Psychiatric Association, D. S. M. T. F. (2013). *Diagnostic and statistical manual of mental disorders : Dsm-5.* Arlington, VA: American Psychiatric Association.

Cook, J. M., Crane, L., & Mandy, W. (2023). Dropping the mask: It takes two. *Autism*.

Fedzechkina, M., Hartley, L. H., & Roberts, G. (2022). Social biases can lead to less communicatively efficient languages. *Language Acquisition*, *30*.

Fedzechkina, M., & 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.

Hull, L., Petrides, K. V., Allison, C., Smith, P., Baron-Cohen, S., Meng, , Lai, C., & Mandy, W. (2017). "putting on my best normal": Social camouflaging in adults with autism spectrum conditions. *Journal of Autism and Developmental Disorders*, *47*, 2519-2534.

Kurumada, C., & Jaeger, T. F. (2015). Communicative efficiency in language production: Optional case-marking in japanese. *Journal of Memory and Language*, *83*, 152-178.

Pearson, A., & Rose, K. (2021). A conceptual analysis of autistic masking: Understanding the narrative of stigma and the illusion of choice. *Autism in Adulthood: Challenges and Management*, *3*, 52.

Roberts, G., & Fedzechkina, M. (2018). Social biases modulate the loss of redundant forms in the cultural evolution of language. *Cognition*, *171C*, 194–201.

Sinnemäki, K. (2008). Complexity trade-offs in core argument marking. 67-88.

Zipf, G. (1935). *The psycho-biology of language* (Vol. ix). Houghton Mifflin.

Feature transmission within concept transmission

Stella Frank[*1] and Serge Belongie[1]

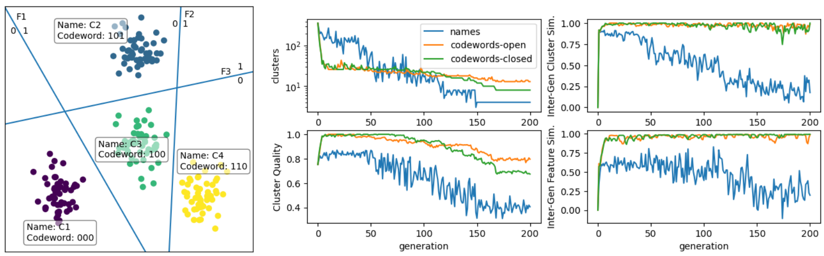
[*]Corresponding Author: stfr@diku.dk
[1]Pioneer Centre for Artificial Intelligence / DIKU, University of Copenhagen, Denmark

Conceptual categories can be described in terms of features: dogs (mostly) have four legs, bark, and shed, while clouds are grey or white and float in the sky. Concepts differ across cultures and languages (Everett, 2013; Majid, 2015), indicating a role for cultural transmission dynamics in their evolution (Contreras Kallens, Dale, & Smaldino, 2018; Carr, Smith, Culbertson, & Kirby, 2020). Since differences in conceptualisations are due to either using different feature boundaries or attending to different feature dimensions (e.g. weather-aware cultures may attend to cloud shape and color in a more fine-grained way), evolving the underlying feature space is integral to concept evolution. In this simulation-based study using Iterated Learning (IL) dynamics (Kirby, 2001), we show how the features underlying concept categories are co-evolved, *given a compositional signalling system* that surfaces the features as well as the concept extensions.

Previous work on concept evolution has focused on discovering concept extensions (Silvey, Kirby, & Smith, 2019; Carr, Smith, Cornish, & Kirby, 2017; Carr et al., 2020) but left the corresponding features implicit. We show that the features themselves, represented as boundaries in high dimensional space, can be also reliably transmitted as part of a compositional concept label.

Model In our framework (Fig. 1a), the world is represented as a high-dimensional *perceptual* feature space; objects are points in this space. *Semantic features* are linear decision boundaries (hyperplanes) in this space, distinguishing points on either side of the boundary. *Concepts* are the interior spaces delimited by the set of features. Borrowing from error correcting output codes (Dietterich & Bakiri, 1995), we represent a concept as a *codeword*, the bitstring representing the feature values corresponding to the concept. A concept also has a *name*, a categorical label. Codewords are by construction *compositional*, while names are *holistic*. While natural languages may not have codeword-like labels (cf. Kirby, Cornish, & Smith, 2008), concepts may be described in terms of their features.

In our IL setup, learner agents infer a semantic feature space, corresponding to a set of concepts, from (label, object) pairs. In the baseline *name* condition, the labels are holistic names, and the task is to learn, via a linear SVM, a hyperplane for



(a) Example Space (2D)

(b) Simulation Results (64D)

Figure 1.: (a) Illustrative example of concept clusters in feature space. (b) Simulation results comparing compositional codewords to holistic names. Synthetic data has 26 well-separated clusters in 64 dimensions. Learners receive 100 labels, in the form of codewords or names, to learn (initally 20) features which they then generalise to 900 unlabeled items. Top left: number of clusters found by each learner (logscale); bottom left: similarity of found clusters to correct clusters, measured using VM (Rosenberg & Hirschberg, 2007); top right: cluster similarity between adjacent generations (learnability) using VM to evaluate the similarity of their labels on a test set; bottom right: feature similarity between adjacent generations, measured as average best-match cosine similarity of feature boundary vectors. Code available at `github.com/scfrank/ecoc_evolang24`.

each name that separates the items with that name from all other items (1-vs-rest). In the *codeword* condition, the labels are codewords composed of feature values. The agent given codewords learns a hyperplane for each feature (e.g., distinguishing items with 0 vs 1 in the nth codeword position), again using a linear SVM. To generate labels for new objects, for the next round of IL, agents use their feature space to determine the conceptual location of a new item (in other words, using the binary features to perform multiclass classification to generate a codeword). This can result in a novel codeword, if this combination of features did not appear in the agent's learning phase. In the 'open world' condition, these new concepts are passed as is to the next generation; in the 'closed world' condition, these novel codewords are mapped to the closest existing codeword using Hamming distance. In the *name* condition, items are always mapped to the closest existing named cluster. In the initial round, names and features are random. At each generation, uninformative features are removed, resulting in shorter codewords.

Results Our simulations (Fig. 1b) show that learning from names alone leads to agents with conceptual systems that are less stable and correspond less to the underlying world, compared to learning from codewords. Codewords also enable IL chains to preserve specificity, and have a natural way of creating new concepts (open-world setting) to counteract the transmission bottleneck.

Acknowledgements

This work was supported in part by the Pioneer Centre for AI, DNRF grant number P1.

References

Carr, J. W., Smith, K., Cornish, H., & Kirby, S. (2017). The Cultural Evolution of Structured Languages in an Open-Ended, Continuous World. *Cognitive Science*, *41*(4), 892–923.

Carr, J. W., Smith, K., Culbertson, J., & Kirby, S. (2020). Simplicity and informativeness in semantic category systems. *Cognition*, *202*, 104289.

Contreras Kallens, P. A., Dale, R., & Smaldino, P. E. (2018). Cultural evolution of categorization. *Cognitive Systems Research*, *52*, 765–774.

Dietterich, T. G., & Bakiri, G. (1995). Solving Multiclass Learning Problems via Error-Correcting Output Codes. *Journal of Artificial Intelligence Research*, *2*, 263–286.

Everett, C. (2013). *Linguistic relativity: Evidence across languages and cognitive domains.* Walter de Gruyer.

Kirby, S. (2001). Spontaneous evolution of linguistic structure - an iterated learning model of the emergence of regularity and irregularity. *5*(2), 102–110.

Kirby, S., Cornish, H., & 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.

Majid, A. (2015). Comparing Lexicons Cross-linguistically. In J. R. Taylor (Ed.), *The Oxford Handbook of the Word.* Oxford University Press.

Rosenberg, A., & Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In J. Eisner (Ed.), *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)* (pp. 410–420). Prague, Czech Republic: Association for Computational Linguistics.

Silvey, C., Kirby, S., & Smith, K. (2019). Communication increases category structure and alignment only when combined with cultural transmission. *Journal of Memory and Language*, *109*, 104051.

Exploring the sound structure of novel vocalizations

Susanne Fuchs[*1], Šárka Kadavá[*1,2,3], Wim Pouw[3], Bradley Walker[4], Nicolas Fay[4], Bodo Winter[5], and Aleksandra Ćwiek[1]

[*]Shared first authorship, corresponding authors: fuchs|kadava@leibniz-zas.de
[1]Leibniz-Zentrum Allgemeine Sprachwissenschaft, Berlin, Germany
[2]Linguistik, Georg-August Universität, Göttingen, Germany
[3]Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands
[4]School of Psychological Science, The University of Western Australia, Perth, Australia
[5]Depart. of English Language and Linguistics, Uni of Birmingham, Birmingham, United Kingdom

When humans speak or animals vocalize, they can produce sounds that are further combined into larger sequences. The flexibility of sound combinations into larger meaningful sequences is one of the hallmarks of human language. To some extent, this has also been found in other species, like chimpanzees and birds. The current study investigates the structure of sounds when speakers are asked to communicate the meaning of 20 selected concepts without using language. Our results show that the structure of sounds between pauses is frequently limited to 1–3 sounds. This structure is less complex than when humans use their native language. The acoustic distance between sounds depends largely on the concept apart from concepts referring to animals, which show a higher diversity of involved sounds. This exploratory analysis might provide evidence of how the structure of sound could have changed from simple to complex in evolution.

1. Introduction

Human speech is composed of small units: sounds that are meaning-distinguishing (phonemes). Several sounds combine into syllables, words, and phrases that carry meaning(s). The sequential order of sounds into larger sequences is a milestone in speech acquisition, and already young infants can start producing sequences of vocalization before they acquire their mother tongue (Wermke, Robb, & Schluter, 2021). Even when language is acquired, nonverbal vocalizations are present in adult communication and are an emerging field of study at the boundaries between non-human and human communication (Pisanski, Bryant, Cornec, Anikin, & Reby, 2022). That means sequences of sounds are not a property of human communication alone but are also found in non-human animals like birds (Sainburg, Theilman, Thielk, & Gentner, 2019; Doupe & Kuhl, 1999; Favaro et al., 2020), meerkats (Rauber, Kranstauber, & Manser, 2020), chimpanzees (Girard-Buttoz et al., 2022). Comparative approaches between human and non-human animal vocalization deserve bottom-up methodologies rather than human-centric analyses (Hoeschele, Wagner, & Mann, 2023). What has been called a syllable in non-human vocalization refers to sound(s) produced between pauses. In human speech production, similar chunks have often been termed inter-pausal units (Bigi & Priego-Valverde, 2019; Prakash & Murthy, 2019). They refer to speech that is realized between pauses.

In this exploratory study, we are interested in sounds realized in novel vocalizations during a charade game, i.e., in a situation where the use of actual words of the participant's language is 'forbidden'. This paradigm has been used to investigate the origin and evolution of language (Fay et al., 2022; Ćwiek et al., 2021; Perlman & Lupyan, 2018).

This paper aims to explore how many sounds are realized between pauses in non-linguistic vocalizations. Furthermore, we investigate the diversity of sounds realized within different concepts, by assessing the distance between them in a multi-variable acoustic space.

2. Methodology

2.1. *Corpus creation*

The present study uses a subset of data collected in a larger study in which participants were recorded performing a series of concepts in three conditions. In the three conditions, participants are asked to communicate a set of concepts using either (1) only gestures, (2) only non-linguistic vocalizations and other sounds, or (3) a combination of gestures and vocalizations. Here, we focus on a subset of the vocalization recordings. We have not analyzed the vocalizations that are produced in the multimodal condition because we assume that first, they are not stand-alone carriers of the meaning, and second, their forms are shaped by the coordination with body motion.

The recordings analyzed here were produced by 62 first-year psychology students at the University of Western Australia (43 female, 17 male, 2 non-binary; aged 17–33, M = 20.21, SD = 3.36). All were speakers of English. Of these, 28 participated in person and 34 remotely via Microsoft Teams, due to COVID-19 restrictions. Participants were allocated 60 concepts to communicate (20 in each modality condition), sampled from a list of 200 concepts comprising the 100-item Leipzig-Jakarta list of basic vocabulary (Tadmor, 2009) plus 100 other basic concepts chosen based on their sensory and modality preferences (Lynott, Connell, Brysbaert, Brand, & Carney, 2020). They were asked to communicate each concept using the specified modality (and without using language) so that another person would be able to view the recording and guess the concept from a list of options. If the participants could not think of a way to communicate a concept, they were permitted to skip it.

2.2. *Concept extraction*

For the exploratory analysis, we focused on a variety of concepts that might reflect different degrees of concreteness and abstraction (see 1). For example, the concept *maybe* is rather abstract or logical than *smoke*. We chose these different concepts to have a wider semantic potential, but have not added categories to the concepts, because a dichotomy between concreteness vs. abstraction has currently been questioned (Banks et al., 2023).

Our analysis only included concepts for which initially at least 5 participants produced vocalizations. For three concepts we excluded acoustic trials as they contained a considerable amount of background noise that made an analysis unreliable.

Table 1. Concepts used in this study. L-J corresponds to the Leipzig-Jakarta list.

Concept	List	No. of speakers
happy	other	6
sad	other	7
bad	other	7
scared	other	5
good	L-J	6
angry	other	7
disgusted	other	7
dog	L-J	6
cat	other	6
bird	L-J	5
fish	L-J	5
fly	L-J	8
old	L-J	4
spoon	other	5
egg	L-J	6
ash	L-J	3
stone/rock	L-J	6
smoke	L-J	4
maybe	other	8
not	L-J	7

2.3. *Acoustic annotation procedures*

The acoustic data were labeled in Praat 6.1.51 (Boersma & Weenink, 2021) by three annotators who are phoneticians by training. Following Swets, Fuchs, Krivokapić, and Petrone (2021), all silent intervals longer than 100 ms were treated as pauses and labeled with ‘p’. Apart from placing boundaries next to pauses, the annotators additionally labeled successive sounds without pauses. The following criteria were used in the decision-making process for separating the speech stream into two or more sounds: a) two (or more) prominent amplitude peaks in the amplitude envelope were present, b) changes in spectral characteristics (e.g., formant structures) were present, and c) sounds were perceptually distinct. Variations in fundamental frequency, e.g., a downward and then upward motion, were only considered as two sounds when they also showed spectral differences in higher frequency ranges and/or differences in the amplitude envelope. All sounds were labeled with an initial ‘s’ and successive numbers when they occurred in a sequence. The first annotator (a1) created the annotation criteria and labeled the data. Annotator 2 (a2) used the available TextGrids from a1 and changed the boundaries when she disagreed. Both agreed on 94.6 percent of the number of sounds. Hereafter, a1 inspected all acoustic files again where disagreement was found and confirmed the changes. Annotator 3 (a3) started labeling from scratch without having TextGrids available. Inter-rater agreement between a2 and a3 was 96.7 percent concerning the overall number of labeled sounds. The temporal differences between the onset of a given sound labeled by a2 and its closed temporal neighbor labeled by a3 were calculated. The same was done for the offset of a sound. The differences were on average 0.048s (median = 0.018s) for the onset and 0.088s (median = 0.027s) for the offset. These differences are influenced by the number of sounds an annotator labeled for a given concept, which makes the calculation of inter-rater agreement challenging. We think that for the current exploratory analysis, the overall agreement is reasonable. We decided to take a2's segmentation for further analysis.

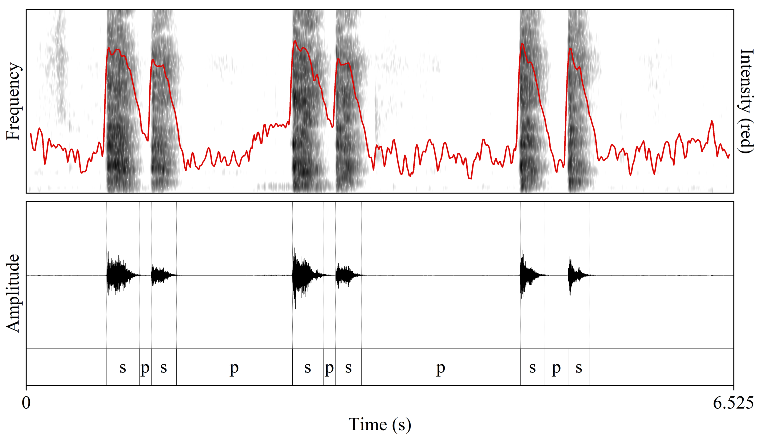


Figure 1. Example for acoustic annotation of the concept *smoke*. All segments are labeled as ‘s’ and pauses as ‘p’. The red line depicts the intensity curve.

2.4. *Analyses of acoustic similarity*

Initially, all audio files were cut into segment-sized files using a custom Python script. Acoustic analysis was performed on these sounds, using the analyze() function of the `soundgen` package in `R` (Anikin, 2019). The output of this function consists of more than 100 acoustic parameters as listed in the documentation (e.g., $f0$, amplitude, formant values, entropy, and their respective mean, median, and standard deviation). Some of these acoustic parameters are present or absent in the recorded sounds, e.g., voicing. However, the presence of voicing is redundant with intensity because voiced sounds are louder than voiceless ones and intensity values can always be calculated. That means, some acoustic parameters are highly correlated and redundant with others. For this reason, we excluded parameters resulting in NA values in the post-processing. Moreover, we excluded voice quality parameters (e.g., flux), because these parameters may have been very sensitive to background noise, which occurred in some speakers. All final parameters were averaged for the whole time series of a sound, and we used mean and standard deviation for further explorations. We ended up with a multidimensional dataset consisting of 45 acoustic parameters. For the analysis of acoustic similarity, we calculated the Euclidean distance between the vector of acoustic parameters of each sound, to all other sounds. As a result, we got a distance matrix that allowed us to extract an average distance between sounds within a trial of a concept and compare it to other concepts.

3. Results and Discussion

3.1. *Structural similarity*

To explore structural similarity, we analyzed if certain sounds occurring between pauses appear alone or in successive order. When speakers try to communicate concepts using novel vocalizations, they frequently realize a relatively small number of sounds between two pauses: 1 sound occurred 208 times, 2 sounds = 80 times, 3 sounds = 35 times, 4 sounds = 24 times, 5 sounds = 11 times, 6 sounds = 3 times, 8 sounds = 4 times, 9 sounds = 1 time, 10 sounds = 1 time, 16 sounds = 2 times, 18 sounds = 1 time. That means structurally most concepts (208 cases in our dataset) are realized with only one sound <s> that is surrounded by pauses. In 80 cases we found realizations of two successive sounds <ss> and in 35 cases participants produced three successive sounds <sss> without being interrupted by a pause. If the data are split by concept, vocalizations for *cat*, *dog*, and *bird* (all within a broader class of animals) also have more than three successive sound combinations, probably mirroring onomatopoeia. For the rest of the data, no conclusions can be drawn, because the number of sounds between pauses is concept-specific.

If pauses are taken into account, sounds were combined flexibly, for example, for four sounds we could get combinations such as <s|s|s|s> or <ss|ss> or <ss|s|s> where | marks a pause.

3.2. *Acoustic similarity*

Similar sounds may be repeated, like in imitating ‘coo-coo’, or they may be of different acoustic quality, like in imitating a cat’s ‘meow’. For this reason, we were further interested in examining the similarity between sounds that make up a novel vocalization.

To have a first look into the diversity of sounds, we analyzed their average acoustic distance within each trial. We preferred this data-driven approach in contrast to labeling the data to phonemic features because it allows us to include sounds that may not occur in the English phoneme inventory, e.g., whistles or clicks. It represents continuous acoustic data instead of putting categorical labels to it, which could also be biased by the native language of the annotator.

Figure 2 depicts the results. We can see that the different concepts vary in their average acoustic distance between sounds. Some abstract concepts like *not* consist of sounds that are closer to each other in distance (i.e., more similar), while *dog* has a larger average acoustic distance between the sounds. Those concepts with several successive sounds (e.g., <sss>) are also the ones with the largest average distance.

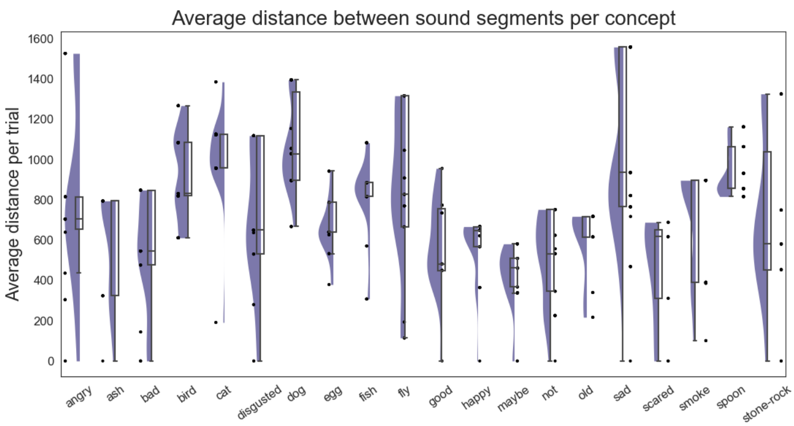


Figure 2. Average acoustic distances between sounds within a single trial displayed by concept, boxplots, and half-violins in purple display data distribution, black dots correspond to single trials. Each concept is displayed at the x-axis and ordered by alphabet.

In summary, the structure of novel vocalizations obtained from a charade game most often contains either one, two, or three successive sounds that are not separated by pauses. This may to some extent be similar to infant's vocalization (Wermke et al., 2021) and non-human species. It is different from human speech production, where already syllables or morphemes can consist of three sounds. Those are combined into larger chunks that are not interrupted by pauses. Our findings suggest that novel vocalizations have a rather simple sound structure that is complexified (i.e., more and probably shorter sounds are realized in a sequence) during language evolution.

4. Supplementary Materials

Dataset and scripts are available on `https://github.com/sarkadava/Evolang2024_SoundSimilarity`.

Acknowledgements

We like to thank the reviewers of Evolang, the participants of the study, and Melissa Ebert for data annotation. This work has been supported by a grant from the German Research Council (FU791/9-1).

References

Anikin, A. (2019). Soundgen: An open-source tool for synthesizing nonverbal vocalizations. *Behavior research methods*, *51*, 778–792.

Banks, B., Borghi, A. M., Fargier, R., Fini, C., Jonauskaite, D., Mazzuca, C., Montalti, M., Villani, C., & Woodin, G. (2023). Consensus paper: Current perspectives on abstract concepts and future research directions. *Journal of Cognition*, *6*(1).

Bigi, B., & Priego-Valverde, B. (2019). Search for inter-pausal units: application to cheese! corpus. In *9th language & technology conference: Human language technologies as a challenge for computer science and linguistics* (pp. 289–293).

Boersma, P., & Weenink, D. (2021). Praat: doing phonetics by computer [computer program](2011). *Version*, *5*(3), 74.

Ćwiek, A., Fuchs, S., Draxler, C., Asu, E. L., Dediu, D., Hiovain, K., Kawahara, S., Koutalidis, S., Krifka, M., Lippus, P., et al.. (2021). Novel vocalizations are understood across cultures. *Scientific Reports*, *11*(1), 10108.

Doupe, A. J., & Kuhl, P. K. (1999). Birdsong and human speech: Common themes and mechanisms. *Annual Review of Neuroscience*, *22*(1), 567–631.

Favaro, L., Gamba, M., Cresta, E., Fumagalli, E., Bandoli, F., Pilenga, C., Isaja, V., Mathevon, N., & Reby, D. (2020). Do penguins' vocal sequences conform to linguistic laws? *Biology letters*, *16*(2), 20190589.

Fay, N., Walker, B., Ellison, T. M., Blundell, Z., De Kleine, N., Garde, M., Lister, C. J., & Goldin-Meadow, S. (2022). Gesture is the primary modality for language creation. *Proceedings of the Royal Society B*, *289*(1970), 20220066.

Girard-Buttoz, C., Zaccarella, E., Bortolato, T., Friederici, A. D., Wittig, R. M., & Crockford, C. (2022). Chimpanzees produce diverse vocal sequences with ordered and recombinatorial properties. *Communications Biology*, *5*(1), 410.

Hoeschele, M., Wagner, B., & Mann, D. C. (2023). Lessons learned in animal acoustic cognition through comparisons with humans. *Animal Cognition*, *26*(1), 97–116.

Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2020). The lancaster sensorimotor norms: multidimensional measures of perceptual and action strength for 40,000 english words. *Behavior Research Methods*, *52*, 1271–1291.

Perlman, M., & Lupyan, G. (2018). People can create iconic vocalizations to communicate various meanings to naïve listeners. *Scientific reports*, *8*(1), 2634.

Pisanski, K., Bryant, G. A., Cornec, C., Anikin, A., & Reby, D. (2022). Form follows function in human nonverbal vocalisations. *Ethology Ecology & Evolution*, *34*(3), 303–321.

Prakash, J. J., & Murthy, H. A. (2019). Analysis of inter-pausal units in indian languages and its application to text-to-speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(10), 1616–1628.

Rauber, R., Kranstauber, B., & Manser, M. B. (2020). Call order within vocal sequences of meerkats contains temporary contextual and individual information. *BMC biology*, *18*, 1–11.

Sainburg, T., Theilman, B., Thielk, M., & Gentner, T. Q. (2019). Parallels in the sequential organization of birdsong and human speech. *Nature communications*, *10*(1), 3636.

Swets, B., Fuchs, S., Krivokapić, J., & Petrone, C. (2021). A cross-linguistic study of individual differences in speech planning. *Frontiers in Psychology*, *12*, 655516.

Tadmor, U. (2009). Loanwords in the world's languages: Findings and results. *Loanwords in the world's languages: A comparative handbook*, *55*, 75.

Wermke, K., Robb, M. P., & Schluter, P. J. (2021). Melody complexity of infants' cry and non-cry vocalisations increases across the first six months. *Scientific reports*, *11*(1), 4137.

Learning Pressures and Inductive Biases in Emergent Communication: Parallels between Humans and Deep Neural Networks

Lukas Galke[*1], Yoav Ram[2], and Limor Raviv[1,3]

[*]Corresponding Author: lukas.galke@mpi.nl
[1]LEADS group, Max Planck Institute for Psycholinguistics, Nijmegen, NL
[2]School of Zoology and Sagol School of Neuroscience, Tel Aviv University, Israel
[3]cSCAN, University of Glasgow, UK

Deep neural networks and humans are two types of learning systems with substantial differences in learning pressures. As many theories of language evolution rely heavily on learning pressures (Kirby et al., 2015; Smith & Kirby, 2008), it is currently unknown whether the learning pressures of humans are sufficiently reflected in deep neural network models in order to allow for insights to carry over and to advance theory building (Dupoux, 2018; Baroni, 2022). In emergent communication simulations, a population of deep neural networks starts from scratch without prior language knowledge and no predefined vocabulary, and are made to develop a language to solve a communication game via reinforcement learning (Lazaridou & Baroni, 2020). While these simulations have great potential for advancing our understanding of the emergence of languages, we can only expect insights gained with deep neural networks to inform language evolution research if the resulting AI languages show similar properties as natural languages (Brandizzi, 2023; Galke et al., 2022). Thus, finding and/or facilitating commonalities (i.e., by introducing appropriate inductive biases) can contribute to our understanding of how languages have evolved.

Reviewing the literature (Galke & Raviv, 2024), we find that the field of emergent communication has successfully designed models to replicate properties of natural languages, even when some of these properties were initially absent in such models. For instance, the lack of a least-effort bias in communicating neural network agents (Chaabouni et al., 2019, 2019; Lian et al., 2023), which gives rise to Zipf's law of abbreviation in natural languages (Kanwal et al., 2017; Zipf, 1949), can be addressed by inducing biases for lazy speakers and impatient listeners (Rita et al., 2020). When going to populations of communicating agents, another case is the absence of a group size effect (Chaabouni et al., 2022), i.e., that larger groups tend to develop more structured languages (Raviv et al., 2019), which can be (to some extent) addressed by introducing variation in learning rates (Rita et al., 2022) or by having agents alternate between sender and receiver roles while restricting parameter updates (Michel et al., 2023). Most importantly, we find that a pressure for learnability, e.g., by having agents continually re-learning the language over and over again – modeled by resetting their parameters (Li & Bowling, 2019; Zhou et al., 2022) – seems to be indispensable for compositional structure to emerge consistently. This pressure for learnability closely resembles the iterated learning paradigm of language evolution research (Smith et al., 2003; Kirby et al., 2014). The necessity of re-learning for structure to emerge is commonly attributed to a learnability advantage of more compositional protocols – or conversely, the ease-of-teaching of compositional protocols to new agents (Li & Bowling, 2019). Although it has been shown for humans (Raviv, de Heer Kloots, & Meyer, 2021), this supposed learnability advantage of compositional structure for learning has not been tested with deep neural networks in a purely supervised learning setting.

Here, we test deep neural networks on their ability to learn new mini-languages with varying degree of compositional structure (Galke, Ram, & Raviv, 2023), analyzing whether more structure leads to more systematic generalization behaviour. We consider long short-term memory models (LSTM) (Hochreiter & Schmidhuber, 1997) trained from scratch as well as a large language model pre-trained on natural language (Brown et al., 2020; Ouyang et al., 2022). We ensure 1:1 comparability to humans by employing the same stimuli and procedure as in a previous study (Raviv et al., 2021). Our results show that – while all languages can be ultimately learned – more systematically structured languages, as quantified by topographic similarity (Brighton & Kirby, 2006), are learned better. Learning more structured languages also leads to more systematic generalizations to new, unseen items, and these generalizations are significantly more consistent and more human-like. Although differences in inductive biases between Transformers and LSTMs need to be taken into account (White & Cotterell, 2021), our findings lead to the clear prediction that children would also benefit from more systematic structure for learning – despite substantial differences in learning patterns compared to adults (Newport, 2020; Hudson Kam & Newport, 2005). This hypothesis is currently being tested (see preregistration: Lammertink et al. (2022)).

In summary, we have shown that deep neural networks display a learning and generalization advantage for more structured and compositional linguistic input – just as (adult) humans. This commonality between humans and machines, combined with other language properties facilitated by inductive biases in emergent communication, provides a rich testbed for using neural networks to simulate the very emergence of language in our species. In ongoing work, we seek to shed new light on why larger populations may tend to develop more structured languages. Notably, this group size effect has been shown to occur in humans even without iterated learning (Raviv et al., 2019), and we hypothesize that modeling cognitive constraints (e.g., memory) would bring us closer towards deep neural networks being useful models for studying human language evolution.

References

Baroni, M. (2022). On the proper role of linguistically oriented deep net analysis in linguistic theorising. In *Algebraic structures in natural language.* CRC Press.

Brandizzi, N. (2023). Toward more human-like ai communication: A review of emergent communication research. *IEEE Access*, *11*, 142317-142340.

Brighton, H., & Kirby, S. (2006). Understanding linguistic evolution by visualizing the emergence of topographic mappings. *Artif. Life*, *12*(2), 229–242.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020).*

Chaabouni, R., Kharitonov, E., Dupoux, E., & Baroni, M. (2019). Anti-efficient encoding in emergent communication. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)* (pp. 6290–6300).

Chaabouni, R., Kharitonov, E., Lazaric, A., Dupoux, E., & Baroni, M. (2019). Word-order Biases in Deep-agent Emergent Communication. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 5166–5175). Florence, Italy: Association for Computational Linguistics.

Chaabouni, R., Strub, F., Altché, F., Tarassov, E., Tallec, C., Davoodi, E., Mathewson, K. W., Tieleman, O., Lazaridou, A., & Piot, B. (2022). Emergent communication at scale. In *Iclr.*

Dupoux, E. (2018). Cognitive Science in the era of Artificial Intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, *173*, 43–59.

Galke, L., Ram, Y., & Raviv, L. (2022). Emergent communication for understanding human language evolution: What's missing? In *Emergent Communication Workshop at the Tenth International Conference on Learning Representations (EmeCom @ ICLR 2022).*

Galke, L., Ram, Y., & Raviv, L. (2023). What makes a language easy to deep-learn? *arXiv preprint abs/2302.12239.*

Galke, L., & Raviv, L. (2024). Emergent communication and learning pressures in language models: a language evolution perspective. *arXiv preprint abs/2403.14427.*

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735-1780.

Hudson Kam, C. L., & 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.

Kanwal, J., Smith, K., Culbertson, J., & 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.

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current opinion in neurobiology*, *28*, 108–114.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102.

Lammertink, I., Bazioni, M., Kloots, M. d. H., & Raviv, L. (2022, July). *Learnability effects in Children: are more structured languages easier to learn?* OSF.

Lazaridou, A., & Baroni, M. (2020). Emergent multi-agent communication in the deep learning era. *arXiv preprint abs/2006.02419*.

Li, F., & Bowling, M. (2019). Ease-of-teaching and language structure from emergent communication. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)* (pp. 15825–15835).

Lian, Y., Bisazza, A., & Verhoef, T. (2023). Communication Drives the Emergence of Language Universals in Neural Agents: Evidence from the Word-order/Case-marking Trade-off. *Transactions of the Association for Computational Linguistics*, *11*, 1033–1047.

Michel, P., Rita, M., Mathewson, K. W., Tieleman, O., & Lazaridou, A. (2023). Revisiting Populations in multi-agent Communication. In *The Eleventh International Conference on Learning Representations (ICLR 2023).*

Newport, E. L. (2020). Children and Adults as Language Learners: Rules, Variation, and Maturational Change. *Topics in Cognitive Science*, *12*(1), 153–169.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *arXiv preprint abs/2203.02155*.

Raviv, L., de Heer Kloots, M., & 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., Meyer, A., & Lev-Ari, S. (2019). Larger communities create more systematic languages. *Proceedings of the Royal Society B*, *286*(1907), 20191262.

Rita, M., Chaabouni, R., & Dupoux, E. (2020). "lazimpa": Lazy and impatient neural agents learn to communicate efficiently. In *Proceedings of the 24th Conference on Computational Natural Language Learning (CoNLL 2020)*

(pp. 335–343). Association for Computational Linguistics.
Rita, M., Strub, F., Grill, J.-B., Pietquin, O., & Dupoux, E. (2022). On the role of population heterogeneity in emergent communication. In *The Tenth International Conference on Learning Representations (ICLR 2022).*
Smith, K., & 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., Kirby, S., & Brighton, H. (2003). Iterated learning: A framework for the emergence of language. *Artificial life*, *9*(4), 371–386.
White, J. C., & Cotterell, R. (2021). Examining the Inductive Bias of Neural Language Models with Artificial Languages. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 454–463). Online: Association for Computational Linguistics.
Zhou, H., Vani, A., Larochelle, H., & Courville, A. (2022). Fortuitous Forgetting in Connectionist Networks. In *The Tenth International Conference on Learning Representations (ICLR 2022).*
Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology.*

The career of verb metaphor: Language evolution parallels online processing differences between nouns and verbs

Dedre Gentner [*1] and Daniel King[1]

[*]Corresponding Author: gentner@northwestern.edu
[1]Department of Psychology, Northwestern University, Evanston, Illinois, US

This paper explores the relationship between online sentence processing and meaning change over time. Specifically, we test the hypothesis, first proposed in Bowdle & Gentner's (2005) *Career of Metaphor*, that novel metaphoric extensions may become new conventionalized word senses over time, driving polysemy. Here we examine whether identified differences between nouns and verbs in online sentence processing—the verb mutability effect—are paralleled by differences in the lexicon, as would be expected if online processing drives lexical changes over time. In Experiment 1, we found that verbs are more polysemous than nouns overall. In Experiment 2, we found that verb senses are rated as being significantly more metaphoric than noun senses, controlling for frequency band; in Experiment 3, we found that historically newer word senses are generally perceived as being more metaphoric than older word senses. Implications for language evolution are discussed.

1. Introduction

Metaphor is widely regarded as an important driver of language change over time (Heine, 1997; Hopper & Traugott, 2003; Xu et al., 2017*)*. One proposal for how this might occur is Bowdle and Gentner's *Career of Metaphor* (CoM) account: that with repeated parallel usage, new figurative uses of words become conventionalized and enter the lexicon as new word senses (Bowdle & Gentner, 2005; Gentner & Bowdle, 2001).[1] Thus, the CoM posits that online novel figurative extensions lead to lexical change over time. This is not a new idea, but empirical evidence linking synchronic and diachronic change is hard to find. Here, we investigate this hypothesis through a novel route. We explore differences between patterns of meaning extension for verbs vs. nouns and trace their consequences for language evolution.

[1] Bowdle and Gentner provided evidence for a further assumption of the CoM theory, namely, *grammatical concordance:* that for noun-noun metaphors, there is a shift in preference from the simile form to the metaphor form with conventionalization.

1.2 Differing patterns of online adjustment between nouns and verbs.

When faced with sentences that show *semantic strain* (e.g., *The lizard worshipped* or *The violin pranced*), people may adjust the standard meaning of one or more word. There is substantial evidence that verbs are more likely to undergo such adjustment than are nouns (Gentner & France, 1988, King & Gentner, 2022; King, 2023). For example, King and Gentner (2022) asked people to paraphrase simple intransitive sentences that varied in degree of semantic strain (e.g., *The husband complained* (low strain) vs. *The motor complained* (higher strain)). Using word2vec (Mikolov et al., 2013), they demonstrated that verbs changed their meanings more under semantic strain than did nouns; and, further, that the *degree* of meaning change increased for verbs (but not nouns) as strain increased. In a further study (Expt. 3), King and Gentner asked raters to judge the type of semantic change that had occurred for the initial noun and verb in each paraphrase. The results showed that verb paraphrases were highly likely to be judged as metaphorically/analogically related to the initial verb. In contrast, noun paraphrases were rarely judged as metaphorically related; rather, they were mostly judged as either taxonomically or metonymically related.[2]

Thus, there are two attested differences—one quantitative and one qualitative—between nouns and verbs in their patterns of online meaning adjustment. First, verbs are more prone to change meaning under semantic strain than are nouns[3] (*verb mutability*; Gentner & France, 1988; King & Gentner, 2022). Second, online verb meaning extensions are more likely to be metaphoric than noun meaning extensions (King & Gentner 2022). If synchronic processes drive diachronic change, these findings predict different patterns of polysemy in the lexicon between nouns and verbs. Here we test three main predictions:

Prediction 1: Verbs should be more polysemous than nouns overall.

Prediction 2: Verb senses in the dictionary should be more metaphoric than noun senses in the dictionary.

Prediction 3: Newer word senses should be more metaphorical, on average, than older word senses. This follows from the CoM prediction that novel metaphoric extensions can become conventionalized over time and enter the lexicon as word senses; with continued usage, these senses will come to be seen as literal.

[2] Here, 'metaphoric' was described as "A term involving an analogy or abstract commonality with the original word"; "Metonymic' was described as 'A term that is associated, rather than similar or taxonomically related (e.g., part-whole) and does not share an abstract commonality"; 'Taxonomic' was described as "superordinate or subordinate."

2. Experiment 1

This study tested the key prediction that verbs will be more polysemous than nouns. A secondary prediction was that high-frequency words will be more polysemous than low-frequency words. Third, we predicted that the effect of frequency on polysemy would be stronger for verbs than for nouns. To do this we obtained polysemy counts for 25,688 nouns and 5,698 verbs.[4]

2.1 Results

To test these predictions, polysemy was modeled as a function of word class (noun vs. verb), word frequency, and the interaction between the two. Thus the design was Frequency X Class. We used an iterative model-comparison approach to select the best-fitting model. A log-transformed second-order exponential model resulted in the best fit: $\log Polysemy \sim (\log Frequency)^2 * Class$. The fitted model was then entered into a Type I ANOVA test of fixed effects. The results bore out all three predictions (Figure 1).

First, as predicted, there was a main effect of Class: verbs had more senses overall (M = 3.25, SD = 3.36) than nouns (M = 2.21, SD = 2.21), $F_{1,\,31380} = 425.14$, $p < .0001$. Second, there was a main effect of word frequency $F_{2,\,31380} = 8377.95$, $p < .0001$. Polysemy increased with frequency for both verbs and nouns, and there was a significant positive exponential relationship between Log(Polysemy) and Log(Frequency). Finally, there was a significant Frequency * Class interaction, $F_{2,\,31380} = 11.672$, $p < .0001$: as predicted, the effect of frequency on polysemy was stronger for verbs than for nouns.

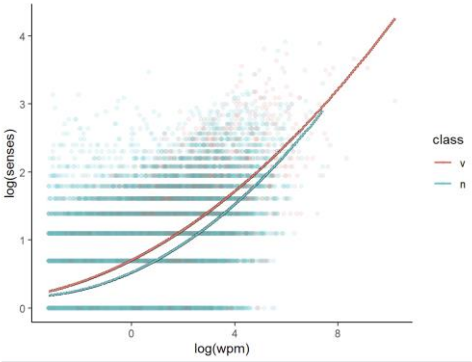


Figure 1. Fitted model results from Experiment 1.

[4] We selected all verbs and nouns from COCA's top 60,000 most frequent English words (lemmatized) Davies, 2008). Polysemy counts for each lemma were obtained by retrieving all senses for every word using the Oxford Dictionary Online API, resulting in 26,888 nouns and 5,750 verbs (32,638 total). Polysemy counts were unavailable for 3,632 lemmas (3,498 nouns and 134 verbs; 11% of the total number), resulting in a net of 31,386 words in the analysis (25,688 nouns and 5,698 verbs).

3. Experiment 2

Experiment 2 tested Prediction 2-–that verb senses in the dictionary will be more metaphoric than noun senses.

3.1 Method

We selected the top 40 most frequent verbs and nouns (lemmatized) from three different frequency bands, determined using Davies (2008): 100 wpm, 10 wpm, and 1 wpm. As in Experiment 1, all senses for each word were obtained using the Oxford Dictionary Online API (1015 senses total).[5] 116 students at a private university in the Midwest served as raters.[6] Participants provided metaphoricity ratings for every sense of each of the 237 lemmas (a total of 1016 senses) using the following procedure.

Each sense was presented to the raters via an example sentence provided by the Oxford Dictionary, with the corresponding lemma bolded: e.g., *The evening had just* ***flown*** *by.* Participants rated the metaphoricity of the bolded word on a 1 to 6 scale. *Metaphoricity* was defined as a word "not being used with its normal literal meaning, but rather with a different meaning that still shares a connection with the normal meaning of the noun."[7] Participants indicated their confidence in each rating on a 1-5 scale and were also able to mark whenever the meaning of the bolded word was unclear in the provided context. Each participant rated only verb senses or only noun senses. This resulted in five ratings per word sense.

The first prediction is that verb senses will be more metaphoric than noun senses overall. Second, we predicted a negative relation between frequency and metaphoricity for both verbs and nouns. This follows from the usage-based conventionalization process proposed in the CoM: the more often a given word sense is used, the more it will be perceived as conventional rather than metaphoric. A final prediction is that high-polysemous words will be rated as more metaphoric than less-polysemous words. On average, if metaphor is a major driver of new sense acquisition, then the more senses a word has, the more metaphoric it should be.

Results

[5] Senses could not be found for three of the selected nouns, leaving a net of 120 verbs and 117 nouns included in the analysis.

[6] All participants answered "yes" to a question asking them if they were native speakers of English.

[7] This definition of 'metaphoricity' is more general that used by King and Gentner (2022), which emphasized analogical relations (see above). In the present case, we wanted to capture any figurative extension, whether analogical or metonymic.

The 116 participants provided a total of 4138 high-confidence ratings.[8] A linear mixed effect model was fit, with metaphoricity rating as the dependent measure, word frequency, word class, their interaction, and word polysemy entered as fixed effects, and subjects and lemma entered as random effects. The fitted model was entered into a Type III ANOVA test of fixed effects using Satterthwaite's method for determining degrees of freedom.

The results supported our two chief hypotheses. First, word senses of verbs were rated as significantly more metaphoric than those of nouns, $F = 4.76$, $p = .03$. Second, for both verbs and nouns, metaphoricity was negatively correlated with word frequency, $F = 9.42$, $p < .0001$. Finally, polysemy was positively related to metaphoricity, $F = 22.16$, $p < .0001$. However, we did not find a significant Frequency * Class interaction. Despite the pattern suggested in Figure 2), the decline in metaphoricity with word frequency was not significantly steeper for verbs than for nouns.

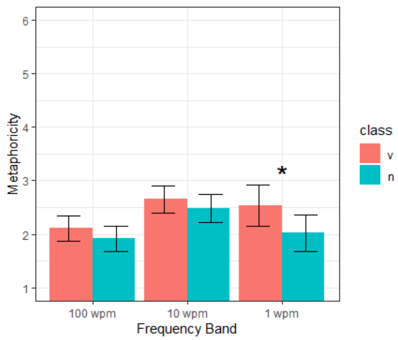


Figure 2. Model results for Experiment 2.

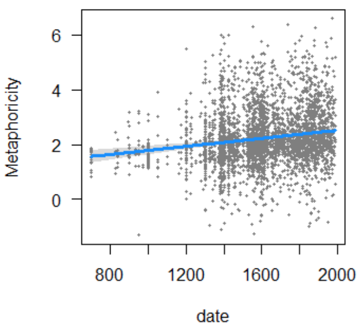


Figure 3. Fitted model results from Experiment 3.

4. Experiment 3

Experiment 3 tested Prediction 3 – that newer senses in the dictionary will be more likely to be labeled as metaphoric than older senses, which have evolved to be seen as literal.

4.1 Method

The 1016 senses for which metaphoricity ratings were obtained in Experiment 2 were used in this experiment. The Oxford Dictionary (OD) API used in that study did not provide the age of the senses queried, so we used the Oxford English Dictionary (OED) online to obtain dates. To match the senses provided

---

[8] There were 5885 ratings in total. We included only high-confidence ratings (4 or 5) and excluded ratings where the participant indicated that the meaning of the word was unclear to them in the context of the sentence, for a net of 4138 ratings included in the analysis.

by the API to those present in the OED, two trained judges, blind to the study's hypotheses, independently identified the closest possible match.[9] For 68 out of the 1016 senses, no date could be obtained; thus, a net of 948 unique senses were included in the analysis.

4.2 Results

A linear mixed effect model was fit, with metaphoricity rating (obtained in Experiment 2) as the dependent measure, sense age as the fixed effect, and subjects (who provided the metaphoricity ratings from Experiment 2) and lemma entered as random effects. As predicted, the effect of sense age was significant, $\beta = 0.1$, $SE = 0.03$, $t = 3.64$, $p < .001$; as the age of the sense decreased, perceived metaphoricity increased (see Figure 3).

5. General Discussion

This research provides novel evidence for the idea that online metaphoric extensions give rise to new word senses, by tracing processing differences between verbs and nouns. There are four main findings. First, verbs are more polysemous than nouns, reflecting the pattern that verbs are more mutable in online sentence understanding. Second, verb senses are more metaphoric than noun senses, reflecting that verbs are more likely to extend metaphorically in online processing than are nouns. Third, word senses for more frequent words are rated as being less metaphorical than those for less frequent words.

Finally, the final experiment directly examined the evolution of word meaning by examining the age of each sense. We found the age of a word's sense predicted the metaphoricity ratings from Experiment 2, such that older senses were rated as more literal than newer senses. This is consistent with the predictions of the Career of Metaphor. Early in its career, a metaphoric sense will be labeled in a dictionary as *figurative*. With continued usage, alignment across uses strengthens the common meaning so that it comes to be seen as *literal* (Bowdle & Gentner, 2005; Gentner & Asmuth, 2019). Further, King and Gentner (2023) found evidence that verb-noun metaphors are processed via a process of structural alignment akin to that used for noun-noun metaphors. Thus it appears that the Career of Metaphor—from novel to conventional meaning—applies to verb metaphors as well as to noun-noun metaphors. Further research may reveal whether this transition occurs more rapidly for verbs than for nouns.

[9] Their judgements were based both on the definitions and the example sentences. Agreement was 70%. In cases where they disagreed, one of the authors made the final decision. When both choices were acceptable, the choice with the earlier date was chosen (57/301 disagreements = 19%).

References

Bowdle, B., & Gentner, D. (2005). The career of metaphor. *Psychological Review, 112*, 193-216.

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. International Journal of Corpus Linguistics, 14(2), 159–190.

Gentner, D., & Asmuth, J. (2019). Metaphoric extension, relational categories, and abstraction. *Language, Cognition and Neuroscience*, *34*(10), 1298-1307.

Gentner, D., & Bowdle, B. F. (2001). Convention, form, and figurative language processing. *Metaphor and Symbol*, 16(34), 223–247.

Gentner, D., Bowdle, B., Wolff, P., & Boronat, C. (2001). Metaphor is like analogy. In D. Gentner, K. J. Holyoak, & B. N. Kokinov (Eds.*), The analogical mind: Perspectives from cognitive science* (pp. 199–253). Cambridge, MA: MIT Press

Gentner, D., & Wolff, P. (1997). Alignment in the processing of metaphor. *Journal of Memory and Language, 37,* 331-355.

Heine, B. (1997). *Cognitive foundations of grammar*. New York: Oxford University Press.

Hopper, P. J., & Traugott, E. C. (2003). *Grammaticalization,* 2nd ed. Cambridge, England: Cambridge University Press.

King, D. (2023). Verb Metaphors as Analogies: Patterns of Meaning Change and the Role of Structure-Mapping in Comprehension [Ph.D., Northwestern University]. https://www.proquest.com/docview/2820900697/abstract/6F793319FE8742A6PQ/1

King, D., & Gentner, D. (2022). Verb metaphoric extension under semantic strain. *Cognitive Science, 46*(5). https://doi.org/10.1111/cogs.13141

King, D., Gentner, D. (2023). Verb metaphors are processed as analogies. In M. Goldwater, F. Anggoro, B. Hayes, & D. Ong (Eds.), *Proceedings of the 45th Annual Meeting of the Cognitive Science Society*.

Xu, Y., Malt, B. C., & Srinivasan, M. (2017). Evolution of word meanings through metaphorical mapping: Systematicity over the past millennium. *Cognitive Psychology*, 96, 41-53.

Zharikov, S., & Gentner, D. (2002). Why do metaphors seem deeper than similes? In W. D. Gray & C. D. Schunn (Eds.*), Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society* (pp. 976–981). Fairfax, VA: George Mason University.

The Mekong-Mamberamo mystery

David Gil[1,], Sihan Chen[2,], and Antonio Benítez-Burraco[*3]

[*]Corresponding Author: abenitez8@us.es
[1]Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
[2]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambrisge, MA, USA
[3]Department of Spanish, Linguistics & Theory of Literature, University of Seville, Spain

This paper suggests that linguistic areas, or sprachbunds, may constitute relics of earlier stages in the evolution of language. Here we focus on the *Mekong-Mamberamo* linguistic area, extending from Mainland Southeast Asia though the Indonesian archipelago and into western New Guinea. The first part of this paper surveys evidence that Mekong-Mamberamo languages exhibit a distinctive grammatical profile associated with greater simplicity in both morphology and syntax. The second part of this paper examines potential explanations for the simple grammatical profile associated with the Mekong-Mamberamo area, and concludes that the most likely of these is that it constitutes an evolutionary relic from an earlier stage in the evolution of language.

1. Introduction

In studying the phylogeny of language, one common method is to try to identify features of contemporary languages that might constitute models for an earlier stage in the evolution of human language. Such evolutionary relics may potentially be present in a number of different domains. First, they may be found embedded in the architecture of particular subsystems of grammar. Thus, Progovac (2015) argues that small clauses, and various other defective clause types, identifiable as part of the more elaborate syntactic structures of languages such as English and Serbo-Croatian, may be viewed as fossils from an earlier evolutionary stage of language. Secondly, certain language types may be viewed more holistically as representative of earlier stages in the evolution of language. Thus, for example, Gil (2017) claims that some contemporary languages, such as Riau Indonesian, come close to instantiating an idealized IMA language — Isolating (lacking internal word structure), Monocategorial (lacking distinct parts of speech), and Associational (lacking construction-specific rules of semantic compositionality) — potentially representing a model for an earlier stage in the evolution of language. Similarly, further down the evolutionary line, Benítez-Burraco and Progovac (2020) suggest that contemporary languages spoken by esoteric, inward-oriented societies are characterized by a more complex morphology alongside a simpler syntax, and that such languages may thus also be considered to represent an earlier stage in the phylogeny of language.

This paper proposes a third domain in which such an evolutionary relic may be observed, namely the linguistic area, or sprachbund. To date, sprachbunds have been mostly used to reconstruct deep stages of language change (e.g. Bickel and Nichols, 2006), but not to infer the types of languages putatively spoken in our remote past. Here we focus on the *Mekong-Mamberamo* linguistic area, first introduced in Gil (2015). This sprachbund consists roughly of mainland Southeast Asia, the Indonesian archipelago, and western parts of the island of New Guinea. Its name derives from the two major rivers located at its two extremities, the Mekong to the north, and the Mamberamo to the east. We survey a body of evidence showing that, compared to a worldwide baseline, Mekong-Mamberamo languages are typically, and sometimes by a substantial margin, associated with greater simplicity with regard to their morphological and syntactic structures. The simple grammatical profile of Mekong-Mamberamo languages poses a mystery: How and why did this profile come into being? This paper suggests that the most likely explanation is that the simple grammatical profile of Mekong-Mamberamo languages is an areally-defined relic representing an earlier stage in the evolution of language.

2. The simple grammatical profile of Mekong-Mamberamo languages

As proposed in Gil (2015), the Mekong-Mamberamo linguistic area is motivated by 17 typical linguistic features. Of these, 7 can be considered as entailing greater grammatical simplicity:

- low differentiation of adnominal attributive constructions
- weakly developed grammatical voice
- isolating word structure
- short words
- low grammatical-morpheme density
- optional thematic-role flagging
- optional Tense-Aspect-Mood (TAM) marking

The Mekong-Mamberamo linguistic area is illustrated with reference to the last of these 7 features, TAM marking, in the following map:

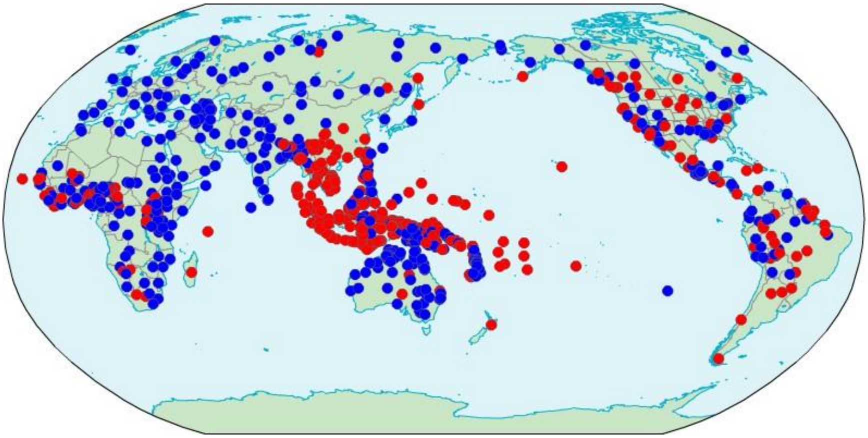


Figure 1. Optional (●) and obligatory (●) Tense-Aspect-Mood marking.

Eyeballing the above map shows clearly that the Mekong-Mamberamo linguistic area, right in the middle of the map, is the only one in which TAM marking is consistently optional. In other parts of the world, such as sub-Saharan Africa, Oceania and the Americas, optional and obligatory TAM-marking languages are interspersed, while in central and western Eurasia and north Africa, obligatory TAM-marking languages are the rule. These geographical patterns are analyzed in more detail in Gil (2021).

Further independent support for the simple grammatical profile of the languages of the Mekong-Mamberamo area derives from a wider study about potential trade-offs between morphology and syntax in the world languages (Benítez-Burraco et al., 2024). In this study, based on the 144 grammatical features listed in WALS, we selected 44 features pertaining to morphological complexity and 39 features pertaining to syntactic complexity. Complexity here is understood in purely descriptive terms: if a grammatical value requires more description than some other vale of the same feature, it is considered as more complex (e.g. Li and Vitányi, 2008; Sinnemäki, 2011). Furthermore, since assigning a grammatical feature to either morphology or syntax can be tricky, and may depend on background theoretical assumptions about the nature of grammar (and even language), we followed the simplest criterion possible: if a grammatical feature pertains to word structure, it was considered as a morphological feature, whereas if it pertains to relationships between words, it was considered as a syntactic feature. To assign each language morphological and syntactic complexity scores, we averaged the normalized values across features pertaining to morphology and syntax respectively. However, due to the limited data availability in WALS, languages vary dramatically in terms of feature coverage. In this study, we considered only the 461 languages in the WALS database for which sufficient data is available.

Figure 2 shows the results of our analysis, plotting the syntactic complexity (S) scores of the 461 languages against the morphological complexity (M) scores, and with the Mekong-Mamberamo languages highlighted in red.

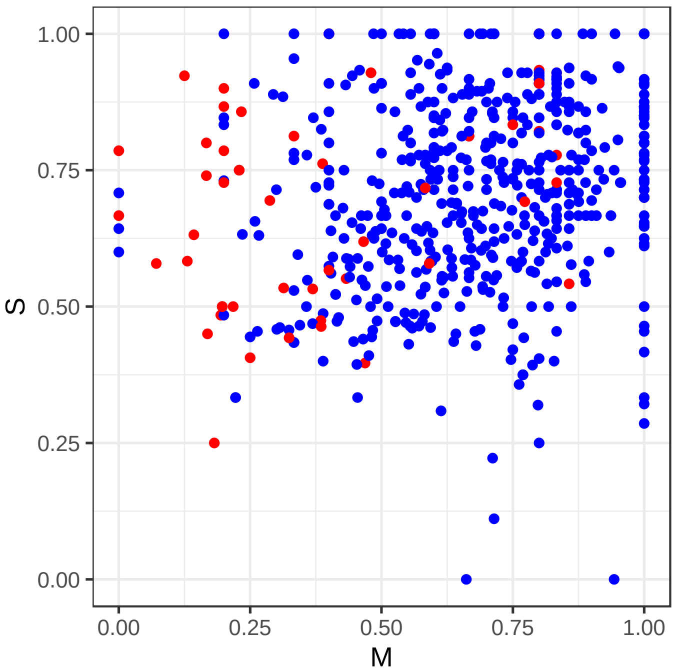


Figure 2. Complexity of Mekong-Mamberamo (●) and other (●) languages. Here the Mekong-Mamberamo area is taken to consist of China south of 30N, all the countries of Mainland Southeast Asia, the Indonesian archipelago (including Timor Este) but not the Philippines, plus New Guinea and associated islands west of 135E.

To evaluate our hypothesis that, following Gil (2015), languages within this area exhibit a simpler grammatical profile, we adopted a Monte-Carlo-based approach, sampling one language from the Mekong-Mamberamo area and another from outside the area and then comparing their morphological complexity and syntactic complexity scores. The 95% confidence interval we obtained was [0.50, 0.67], significantly higher than 0.25 (the threshold for a by-chance relationship). Overall, the results supported our hypothesis.

3. Towards a resolution of the mystery

The Mekong-Mamberamo linguistic area thus presents us with a mystery: How could a large geographical region have come to be associated with a grammatical profile so systematically different (and simpler) from that found in other parts of the world? In Gil (2015:412-4), a number of speculative answers are put forward, appealing to factors such as language contact and the relatively recent presence of other hominin species; however, none of these answers enjoys clear cut support. In lieu of such support, our default hypothesis is that the Mekong-Mamberamo linguistic area is the way that it is because it always was like that. In other words: the simple grammatical profile characteristic of Mekong-Mamberamo languages is a relic from an earlier stage in the evolution of language in which languages exhibited less complex morphologies and syntaxes.

As is well known, the diachronic accretion of complexity is a ubiquitous process, observable throughout the world even in relatively shallow time depths. In view of this, one may indeed wonder how likely it is that a geographical region as large as the Mekong-Mamberamo might have been spared such pervasive diachronic processes of complexification for such a long time, allowing for the preservation of an earlier stage in the evolution of language itself. Consideration of contact between closely related dialects or languages reveals some of the mechanisms that might have contributed to the perseverance, over time, of the simpler Mekong-Mamberamo grammatical profile (Gil 2020:190-1). For example, in dialects of Kerinci spoken in central Sumatra, most words occur in one of two competing forms, absolute and oblique; however, in the emerging Kerinci koiné, the absolute/oblique alternation is in the process of disappearing, under influence from surrounding Mekong-Mamberamo languages. Processes such as these suggest that the simpler grammatical profile associated with Mekong-Mamberamo languages may be self-perpetuating, providing a second pole of stability around which languages may cluster and persevere.

An apparent challenge to the archaic nature of the simple Mekong-Mamberamo profile derives from the diachronic study of the Austronesian languages occupying a large central swathe of the area. It is commonly accepted that the original grammatical profile of Proto-Austronesian was the more complex one that is currently observable in contemporary Austronesian languages of Taiwan and the Philippines, and that the simpler grammars of many of the other Austronesian languages of the Indonesian archipelago are due to contact-induced simplification that took place some 3,500-4,000 years ago, when Austronesian languages spread south into the archipelago. According to this view, their simpler grammatical profiles would actually be an innovation dating back just a few thousand years at the most. However, in Gil (2020) it is argued that the Austronesian languages of the Indonesian archipelago are most appropriately viewed as exhibiting dual heritage, reflecting the coming together of two distinct linguistic lineages: while the vocabulary is largely Austronesian, much of the grammar represents a direct inheritance from the non-Austronesian languages that were already present in the region. In other words, the simpler grammatical profiles associated with today's Austronesian languages of the Indonesian archipelago predate the arrival of Austronesian languages in the region; their presence in the region was a continuous one, dating back as far as we can see.

Clearly, at an earlier point in human evolution, languages were simpler than they are today. Accordingly, the plausibility of the hypothesis that the Mekong-Mamberamo linguistic profile is an evolutionary relic depends on how far back in human pre-history one must go until all the world's languages exhibited the simpler grammatical profiles of today's Mekong-Mamberamo languages. Consideration of the worldwide geographical distribution of grammatical complexity suggests that this might have been at a relatively recent stage, post-dating the spread of modern humans out of Africa. As argued in Gil (2009), a simple IMA language is all that was needed to facilitate collective tasks such as sailing a boat to an island over the horizon. Thus, humans could have spread out all over the world, speaking languages associated with a simple grammatical profile resembling that of contemporary Mekong-Mamberamo languages. Later, complexification would have occurred, arising independently in several locations, and then spreading until it encompassed most of the world — with the exception of the Mekong-Mamberamo area. In the same vein, Benítez-Burraco and Progovac have hypothesized that humans spoke simpler languages perhaps as late as 50.000 years ago, at which time languages began to complexify under the effects of our increased prosocial behavior. All in all, in accordance with such scenarios, the Mekong-Mamberamo linguistic area may have been left behind as a relic of an earlier stage in the evolution of language. It remains to be determined which specific factors, seemingly extralinguistic by nature (social, cultural, or even environmental), contributed to preserve these relic features in this part of the world.

References

Bickel, B., & Johanna (2006) "Oceania, the Pacific Rim, and the theory of linguistic areas". In *Annual Meeting of the Berkeley Linguistics Society*, 32(2), 3-15.

Benítez-Burraco, Antonio and Ljiljana Progovac (2020) "A four-stage model for language evolution under the effects of human self-domestication", *Language and Communication 73.1–17*

Chen, Sihan, David Gil, Sergey Gaponov, Jana Reifegeste, Tessa Yuditha, Tatiana Tatarinova, Ljiljana Progovac and Antonio Benitez-Burraco (2023) "Linguistic and Memory Correlates of Societal Variation, A Quantitative Analysis", PsyArXiv Preprints, 10.31234/osf.io/bnz2s.

Gil, David (2009) "How Much Grammar Does It Take to Sail a Boat?", in G. Sampson, D. Gil and P. Trudgill eds., *Language Complexity as an Evolving Variable*, Oxford University Press, Oxford, 19-33.

Gil, David (2015) "The Mekong-Mamberamo Linguistic Area", in N.J. Enfield and B. Comrie eds., *Languages of Mainland Southeast Asia, The State of the Art*, Pacific Linguistics, DeGruyter Mouton, Berlin, 266-355.

Gil, David (2017) "Isolating-Monocategorial-Associational Language", in H. Cohen and C. Lefebvre eds, Handbook of Categorization in Cognitive Science, Second Edition, Elsevier, Amsterdam, 471-510.

Gil, David (2020) "Dual Heritage: The Story of Riau Indonesian", in D. Gil and A. Schapper eds., *Austronesian Undressed: How and Why Languages Become Isolating*, John Benjamins, Amsterdam, 119-212.

Gil, David (2021) "Tense-Aspect-Mood Marking, Language Family Size, and the Evolution of Predication", in A. Benítez-Burraco and L. Progovac eds., Prehistorical Languages, *Philosophical Transactions B.*

Progovac, Ljiljana (2015). *Evolutionary Syntax*. Oxford: Oxford University Press.

Using Genetics to Investigate the Evolution of Language and Speech: New Findings on Musicality and Vocal Acoustics

Rósa S. Gísladóttir[*1,2] and Kári Stefánsson[1,3]

[*]Corresponding Author: rosas@hi.is
[1] deCODE Genetics/Amgen Inc., Reykjavik, Iceland
[2] Department of Icelandic and Comparative Cultural Studies, University of Iceland, Reykjavik, Iceland
[3] Faculty of Medicine, University of Iceland, Reykjavik, Iceland

Genetic data have for over 20 years been recognized as one of the most promising avenues for empirical language evolution research (Christiansen & Kirby, 2003; Fisher, 2017; Fitch, 2017). Genes have been described as the closest thing to "fossils of language", with archaic human DNA akin to a "time machine" (Fitch, 2017). The assumption is that genome-wide association studies (GWAS) and other genetic research can pinpoint genes involved in speech and language, providing a springboard for subsequent analyses of primate and archaic human genomes that can shed light on the timeline for language evolution in our ancestors (Christiansen & Kirby, 2003; Fisher, 2017; Fitch, 2017). GWAS on speech and language has been hampered by a lack of large cohorts with genetic information and relevant phenotypes, although important advances have recently been made (Doust et al., 2022; Eising et al., 2022). Here we discuss new results on the genetics of speech acoustics and musicality traits. Our aim is to illustrate the different ways in which genetic research can test and inform theorizing on the evolution of language and speech broadly construed.

We begin with a new attempt at the "time machine" strategy. To better understand genetic factors influencing speech acoustics, we performed a GWAS on voice pitch (f0) and vowel formants in a population with limited dialectal differences (N = 12,901) (Gisladottir et al., 2023). We discovered sequence variants in *ABCC9* that influence voice pitch and other traits, including pulse pressure and the expression of *ABCC9* in the adrenal gland (of potential relevance for proposals linking vocal behavior with self-domestication and adrenal gland function; Benítez-Burraco et al., 2018; Ghazanfar et al., 2020; Wilkins et al., 2014). Since the vocal channel plays a relatively more important role in humans than in other great apes (Corballis, 2002; Levinson & Holler, 2014), we compared the human *ABCC9* to other primates, identifying a missense change in *ABCC9* that is fixed in humans but not present in primate reference genomes. When did this missense change emerge? By examining four genomes from archaic humans, we conclude that the missense change occurred after hominins split from the great apes but before they diverged into modern humans, Neanderthals, and Denisovans. The implications of this finding are far from clear. Voice pitch is a simple acoustic measure without direct relevance for vocal learning or cooperative behavior. Nevertheless, this study is a reminder that the more we know about the genetic components involved in human communication at all levels, the better we will be able to sketch how speech and language evolved in our ancestors.

There are several limitations of the strategy above. A single gene account for a trait is implausible, given the messy mappings between genetics and complex traits (Fisher & Vernes, 2015). However, there are ways to leverage the general genetic architecture behind a trait, which we illustrate with a study on human musicality. Since Darwin, several authors have proposed that the origins of language can be traced to a musical or prosodic proto-language, with the evolution of vocal imitation for singing as a key stepping stone (Darwin, 1871; Fitch, 2010). Fitch has pointed out some testable predictions that emerge from this account, noting that "because the neural mechanisms underlying song were precursors of phonological mechanisms in spoken language, we expect considerable overlap between phonological and musical abilities (within individuals) and mechanisms (across individuals)," (Fitch, 2010, p. 506). To test this prediction, we performed a GWAS of musicality traits, using tests of musical pitch and beat perception (Peretz & Vuvan, 2017) and self-reported music perception and training (Müllensiefen et al., 2014) (N = 20,440, age 18-95 years). We found that musicality traits correlate with speech and language traits at the phenotypic level. To test overlap of the genetic mechanisms, we then estimated the genetic correlation of the musicality traits with 26 other cognitive traits. Besides genetic correlations with intelligence and personality for some measures, we found that all musicality traits show substantial genetic correlation with verbal working memory, also known as the phonological loop ($rg = 0.43$ to 0.30, $P < 1.3\times10^{-5}$). Verbal working memory is necessary to learn complex utterances and thus relevant for vocal learning (Aboitiz, 2018). While the causal scenarios underlying genetic correlations are difficult to entangle, these findings are in line with the view that musicality and spoken language share genetic roots.

Finally, we will turn back to the GWAS on speech acoustics. We estimated the heritability of voice pitch and vowel formants, providing an estimate of phenotypic variance explained by common sequence variants (SNP-based heritability). We discovered that even vowel formants have a small-to-modest SNP-based heritability, particularly F_2 (14%). This finding has bearing on the proposal that genetic biases influencing the vocal tract can be amplified through language transmission, ultimately contributing to linguistic diversity (Dediu et al., 2017, 2019).

Each of these strategies discussed above brings numerous complexities. However, the promise of GWAS for language evolution remains tantalizing, and it is now more attainable than ever due to fast developments in population genomics.

References

Aboitiz, F. (2018). A brain for speech. Evolutionary continuity in primate and human auditory-vocal processing. *Frontiers in Neuroscience*, *12*, 174. https://doi.org/10.3389/fnins.2018.00174

Benítez-Burraco, A., Theofanopoulou, C., & Boeckx, C. (2018). Globularization and domestication. *Topoi*, *37*(2), 265–278.

Christiansen, M. H., & Kirby, S. (2003). Language evolution: Consensus and controversies. *Trends in Cognitive Sciences*, *7*(7), 300–307.

Corballis, M. C. (2002). *From hand to mouth: The origins of language*. Princeton University Press.

Darwin, C. (1871). *The descent of man, and selection in relation to sex*. Murray.

Dediu, D., Janssen, R., & Moisik, S. R. (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., & Moisik, S. R. (2019). Weak biases emerging from vocal tract anatomy shape the repeated transmission of vowels. *Nature Human Behaviour*, *3*(10), 1107–1115.

Doust, C., Fontanillas, P., Eising, E., Gordon, S. D., Wang, Z., Alagöz, G., Molz, B., & Pourcain, B. S. (2022). Discovery of 42 genome-wide significant loci associated with dyslexia. *Nature Genetics*, *54*(11), 1621–1629.

Eising, E., Mirza-Schreiber, N., de Zeeuw, E. L., Wang, C. A., Truong, D. T., Allegrini, A. G., Shapland, C. Y., Zhu, G., Wigg, K. G., Gerritse, M. L., Molz, B., Alagöz, G., Gialluisi, A., Abbondanza, F., Rimfeld, K., van Donkelaar, M., Liao, Z., Jansen, P. R., Andlauer, T. F. M., … Fisher, S. E. (2022). Genome-wide analyses of individual differences in quantitatively assessed reading- and language-related skills in up to 34,000 people. *Proceedings of the National Academy of Sciences*, *119*(35), e2202764119. https://doi.org/10.1073/pnas.2202764119

Fisher, S. E. (2017). Evolution of language: Lessons from the genome. *Psychonomic Bulletin & Review*, *24*(1), 34–40. https://doi.org/10.3758/s13423-016-1112-8

Fisher, S. E., & Vernes, S. C. (2015). Genetics and the language sciences. *Annual Review of Linguistics*, *1*(1), 289–310.

Fitch, W. T. (2010). *The evolution of language*. 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

Ghazanfar, A. A., Kelly, L. M., Takahashi, D. Y., Winters, S., Terrett, R., & Higham, J. P. (2020). Domestication phenotype linked to vocal behavior in marmoset monkeys. *Current Biology*, *30*(24), 5026–5032.

Gisladottir, R. S., Helgason, A., Halldórsson, B. V., Helgason, H., Borsky, M., Chien, Y.-R., Gudnason, J., Gudjonsson, S. A., Moisik, S., Dediu, D., Thorleifsson, G., Tragante, V., Bustamante, M., Jónsdóttir, G., Stefánsdóttir, L., Magnússon, S. H., Harðarsson, M., Ferkingstad, E., Halldorsson, G. H., … Stefansson, K. (2023). Sequence variants affecting voice pitch in humans. *Science Advances*, *9*(23), eabq2969.

Levinson, S. C., & 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

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The Musicality of Non-Musicians: An Index for Assessing Musical Sophistication in the General Population. *PLOS ONE*, *9*(2), e89642. https://doi.org/10.1371/journal.pone.0089642

Peretz, I., & Vuvan, D. T. (2017). Prevalence of congenital amusia. *European Journal of Human Genetics*, *25*(5), 625–630.

Wilkins, A. S., Wrangham, R. W., & Fitch, W. T. (2014). The “Domestication Syndrome” in mammals: A unified explanation based on neural crest cell behavior and genetics. *Genetics*, *197*(3), 795–808. https://doi.org/10.1534/genetics.114.165423

An event-based model for linguistic phylogenetics

David Goldstein*[1], Shawn McCreight[2], Éva Buchi[3], and John Huelsenbeck[4]

*Corresponding Author: goldsteindm@gmail.com
[1]Department of Linguistics, University of California, Los Angeles, Los Angeles, USA
[2]Nytril LLC, Gig Harbor, USA
[3]Centre national de la recherche scientifique, Paris, France
[3]Université de Lorraine, Nancy, France
[4]Department of Integrative Biology, University of California, Berkeley, Berkeley, USA

Introduction. In linguistic phylogenetics inferences are standardly drawn from lexical cognate relationships, which are represented with abstract discrete values such as 0 and 1 in the case of binary characters (e.g., Bouckaert et al., 2012; Greenhill & Gray, 2012; Chang, Cathcart, Hall, & Garrett, 2015). Despite the prevalence of this approach, it suffers from well-known flaws.

Table 1. Cognate word-forms in Romance for 'stone'

Language	Aligned cognate word-forms					
Latin	p		e	t	r	a	m
Portuguese	p		ɛ	ð	ɾ	a	
Spanish	p	j	e	d	ɾ	a	
Catalan	p		e	d	ɾ	ə	
French	p	j	ɛ		ʁ		
Italian	p	j	ɛ	t	r	a	
Romanian	p	j	a	t	r	ə	

First, it discards a massive amount of information. Consider the Romance word-forms in Table 1, which all descend from a common ancestor. Under the conventional approach, they would all be assigned to the same cognate class. Although identical in this respect, they have diverged segmentally. It is precisely this segmental divergence that the standard practice ignores. Second, the representation of cognate relationships relies on arbitrary values, which lack consistent reference across cognate sets (Wright, Lloyd, & Hillis, 2016, 602). As a result, the standard approach does not model events of lexical change directly and estimated transition rates are not linguistically meaningful.

Incorporating segmental information. The TKF91 model overcomes these problems by modeling segmental changes among cognate word-forms (Thorne, Kishino, & Felsenstein, 1991; Lunter, Miklós, Song, & Hein, 2003). Under this model, one of three events is possible in an instant of time: an insertion of a single segment, a deletion of a single segment, or a transition from one segment to another. These are the very processes that give rise to the Romance word-forms in Table 1. Insertions and deletions are modeled as continuous-time birth-death processes, while substitution models such as JC69 or GTR are used for transitions between segments. This talk presents the first application of the TKF91 model to linguistic data.

Data and methods. Parameters are estimated in a Bayesian-MCMC framework, with estimates based on aligned phonemic sequences of 2,628 cognate word-forms from 9 Romance languages and Latin. Concepts for the cognate sets are selected from the Swadesh 207-word list. The model is provided with initial alignments, but they are marginalized over, so posterior distributions are not conditioned on any particular one. Tree topologies and branch lengths can also be estimated in this framework, but here I focus on transition rates.

Results and Discussion. Estimates of segmental volatility are presented in Figure 1. Vowels are on the whole more volatile than consonants, with long vowels and diphthongs being particularly unstable. Transition rates within each segmental class are remarkably similar.

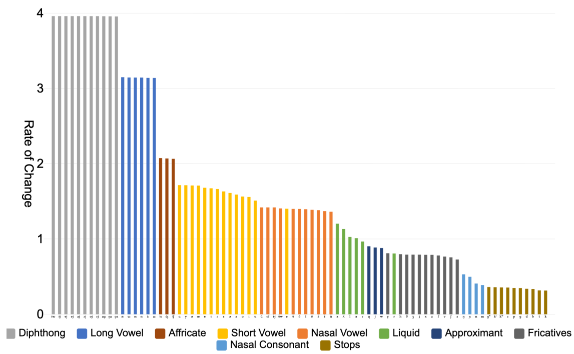


Figure 1. Segmental volatility

The event-based approach of the TKF91 model offers significant benefits. First, it allows scholars to take advantage of the rich information in words when drawing phylogenetic inferences. Second, it has enormous potential for phonology, since it provides the first phylogenetically based method for estimating the evolutionary stability of phonemes and phonetic segments. More broadly, the TKF91 model brings linguistic phylogenetics closer to the study of molecular phylogenetics, in as much as segmental sequences parallel those of nucleotides.

References

Bouckaert, R. R., Lemey, P., Dunn, M., Greenhill, S. J., Alekseyenko, A. V., Drummond, A. J., Gray, R. D., Suchard, M. A., & Atkinson, Q. D. (2012). Mapping the origins and expansion of the Indo-European language family. *Science*, *337*(6097), 957–960.

Chang, W., Cathcart, C. A., Hall, D. P., & Garrett, A. J. (2015). Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language*, *91*(1), 194–244.

Greenhill, S. J., & Gray, R. D. (2012). Basic vocabulary and Bayesian phylolinguistics. *Diachronica*, *29*(4), 523–537.

Lunter, G. A., Miklós, I., Song, Y. S., & Hein, J. (2003). An efficient algorithm for statistical multiple alignment on arbitrary phylogenetic trees. *Journal of Computational Biology*, *10*(6), 869–889.

Thorne, J. L., Kishino, H., & Felsenstein, J. (1991). An evolutionary model for maximum likelihood alignment of DNA sequences. *Journal of Molecular Evolution*, *33*(2), 114–124.

Wright, A. M., Lloyd, G. T., & Hillis, D. M. (2016). Modeling character change heterogeneity in phylogenetic analyses of morphology through the use of priors. *Systematic Biology*, *65*, 602–611.

The iconic affordances of gesture and vocalization in emerging languages in the lab

Oxana Grosseck[*1,2], Marcus Perlman[3], Gerardo Ortega[3], and Limor Raviv[2,4]

[*]Corresponding Author: oxana.grosseck@mpi.nl
[1]Radboud University, Nijmegen, Netherlands
[2]LEADS, Max Planck Institute for psycholinguistics, Nijmegen, Netherlands
[3]English Language and Linguistics, University of Birmingham, UK
[4]cSCAN, University of Glasgow, Glasgow, UK

Iconic signals play an essential role in bootstrapping a novel communication system and getting it off the ground (e.g., Fay et al., 2013; Perlman et al, 2015). However, iconicity may not be uniformly or readily available across modalities, and the relative iconic affordance of speech vs. gestures has long been a subject of interest and controversy given its relevance to language origins (e.g., Kendon, 2017). Specifically, while there is a wide consensus that the gestural modality holds great potential for iconicity (which is often taken as support for the gesture-first hypothesis for language evolution, e.g., Fay et al, 2014; Corballis, 2002), recent work suggests that the vocal modality affords more iconicity than previously thought (e.g. Perlman, 2017; Dingemanse et al. 2015) - supporting a more multimodal view of language origins. Yet despite the importance of assessing the relative iconic affordance of each modality during the emergence of a novel communication system, only a handful of studies directly compared the communicative success of novel signal creation across modalities using the same stimuli and experimental procedure (Macuch-Silva et al., 2020; Fay et al., 2013, 2014, 2022; Lister et al., 2021). Of these, only two (Lister et al., 2021, Fay et al., 2022) examined signals' degree of iconicity by measuring guessing accuracy with naïve participants - but only for known concepts and signals produced in isolation (i.e., not during communication or as a part of a structured system). Furthermore, it is still unclear how iconicity evolves over time during the formation of a novel communication system. While some work suggests that iconicity decreases over repeated interactions to make space for more systematic and/or efficient signals (tones: Verhoef et al., 2016; drawings: Fay & Ellison, 2013), the one study that tested this with vocalizations found that iconicity *increased* over rounds, alongside conventionalization (Perlman et al., 2015) – suggesting that iconicity trajectories may be modality-specific.

Here, we present the first empirical study to directly test the iconic affordances of the gestural and vocal modalities for different dimensions of meaning (shape, size, motion, speed), as well as how this iconicity changes over time across modalities during the formation of a new language. To this end, we introduce a novel paradigm for evaluating the fine-grained iconicity of productions with respect to individual referent features (Fig. 1). In a pre-registered online experiment (https://osf.io/gh6xp), >1200 naïve participants are exposed to audio/video recordings of one vocal and one gestural sign referring to novel multi-dimensional stimuli. These were collected from 18 dyads playing an emergent referential communication game for multiple rounds in a virtual environment (Motiekaityte et al., in prep). Upon exposure, participants are asked to guess the meanings depicted in the recording following a 4-step decision tree, with each step corresponding to one feature of the referent, and with the alternatives at every step being determined by the participant's previous choice. For each feature, we record guessing accuracy as well as whether the participant believed the feature was encoded in the video.

We predict that: (H1) iconicity will be present in both gestures and vocalizations, yet with overall more iconicity in gestures; (H2) Iconicity trajectories over the course of dyadic interaction will differ across modalities, with iconicity decreasing for gestures (i.e., lower guessing accuracy for later productions) and possibly increasing for vocalizations; (H3) Some features (e.g., shape) will be better guessed in gestures, while others (e.g., speed) will be guessed well across modalities. Based on preliminary results from N=300 participants, gestures are indeed guessed better than vocalizations (H1), especially for features like shape and motion (H3).

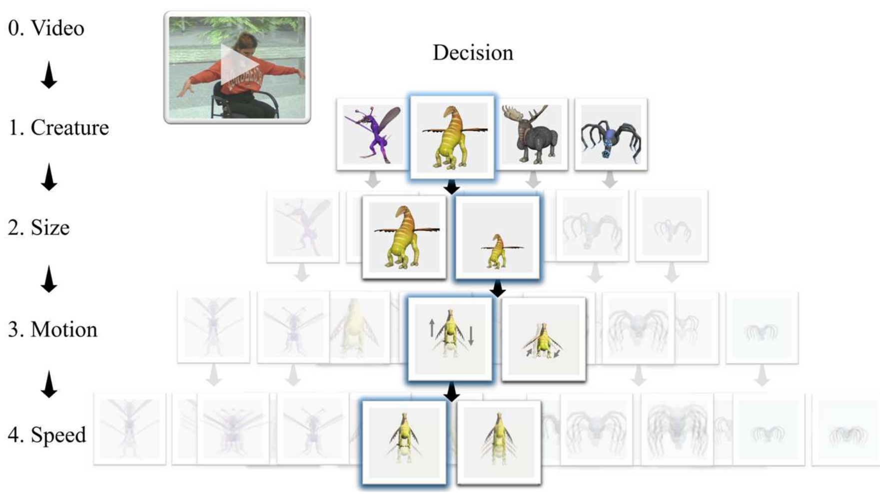


Figure 1. Design of the iconicity experiment, with an example trajectory through the decision tree.

References

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, iconicity, and systematicity in language. *Trends in Cognitive Sciences, 19*(10), 603–615.

Corballis, M. C. (2002). *From hand to mouth: The origins of language.* Princeton University Press.

Fay, N., Arbib, M., & Garrod, S. (2013). How to bootstrap a human communication system. *Cognitive science, 37*(7), 1356-1367.

Fay, N., Ellison, T.M. (2013). The Cultural Evolution of Human Communication Systems in Different Sized Populations: Usability Trumps Learnability. *PLoS ONE, 8*(8): e71781.

Fay, N., Lister, C. J., Ellison, T. M., & Goldin-Meadow, S. (2014). Creating a communication system from scratch: gesture beats vocalization hands down. *Frontiers in Psychology*, *5*, 354.

Fay, N., Ellison, M., & Garrod, S. (2014). Iconicity: From sign to system in human communication and language. *Pragmatics & Cognition*, *22*(2), 244-263.

Fay, N., Walker, B., Ellison, T. M., Blundell, Z., De Kleine, N., Garde, M., ... & Goldin-Meadow, S. (2022). Gesture is the primary modality for language creation. *Proceedings of the Royal Society B*, *289*(1970), 20220066.

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

Lister, C. J., Burtenshaw, T., Walker, B., Ohan, J. L., & Fay, N. (2021). A Cross-Sectional Test of Sign Creation by Children in the Gesture and Vocal Modalities. *Child Development*, *92*(6), 2395-2412.

Macuch Silva. V., Holler, J., Ozyurek, A., Roberts, S.G. (2020). Multimodality and the origin of a novel communication system in face-to-face interaction. *R. Soc. open sci. 7*: 182056.

McLean, B., Dunn, M., & Dingemanse, M. (2023). Two measures are better than one: Combining iconicity ratings and guessing experiments for a more nuanced picture of iconicity in the lexicon. *Language and Cognition,* 1-24.

Perlman, M., Dale, R., & Lupyan, G. (2015). Iconicity Can Ground the Creation of Vocal Symbols. *Royal Society Open Science 2* (8): 150152.

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.

Verhoef, T., Kirby, S., & De Boer, B. (2016). Iconicity and the emergence of combinatorial structure in language. *Cognitive science*, *40*(8), 1969-1994

Grammar change through cultural transmission

Juan Guerrero Montero*[1], Richard A. Blythe[1], Simon Kirby[2], Dan Lassiter[2], and Robert Truswell[2]

*Corresponding Author: j.a.guerrero-montero@sms.ed.ac.uk
[1]School of Physics and Astronomy, University of Edinburgh, Edinburgh, United Kingdom
[2] School of Philosophy, Psychology, and Language Sciences, University of Edinburgh, Edinburgh, United Kingdom

The inter-generational transmission of language, which underpins language evolution, has long been modelled as a process of Iterated Bayesian Learning (IBL; Griffiths and Kalish, 2007). These models involve agents producing linguistic data in the form of utterances that agents in the next generation use to learn the language. This learning process involves combining this linguistic input with a prior distribution representing their inductive biases. The IBL paradigm has been proven equivalent to the Wright-Fisher (WF) model from population genetics (Reali & Griffiths, 2010), which in turn provides access to quantitative tools for analysing language change in corpus data (Newberry et al., 2017).

Nevertheless, IBL has qualities that do not reflect those of natural languages. First, its emerging stationary distribution over languages depends only on the inductive biases contained in the prior and not on the communication process, assuming that speakers sample from their posterior. Secondly, this stationary distribution respects detailed balance, implying that evolution processes are equally likely to happen in the forward and backwards directions. However, language change is directional, as evidenced by e.g. the irreversibility of grammaticalisation (Haspelmath, 1999). While certain extensions of IBL avoid these issues, e.g. by allowing learning from several agents (Smith, 2009) or changing the production strategy (Kirby, Dowman, & Griffiths, 2007), their mathematical complexity usually weakens the equivalence of the model to WF. This hinders application to corpus analysis, as the WF model is efficient, well-studied, and can be customised to include a wide variety of evolutionary effects.

Here, we introduce a model of grammar change where F linguistic functions represented by E expressions co-evolve following an IBL paradigm. In it, grammars are composed of probabilities g_{fe} representing speakers' expectation that expression e is used to express function f. While the model can include effects like production errors and biases, its key component is the inclusion of imperfect understanding, implemented as a probability that the learner infers function f' when f was meant. Figure 1, top, represents the model schematically.

We show that imperfect understanding breaks convergence to the prior and detailed balance, thus accounting for the directionality of language change as well as communicative effects on the statistical properties of language, while remaining equivalent to a set of F co-evolving WF processes (one per function). This feature allows us to quantify and discriminate between effects like drift, understanding and production errors, analogy and social selection, which are all parametrised differently in the WF paradigm. This enables a full analysis of grammar change in corpus data.

We apply the model to data on the diachronic use of relativisers in Middle and Modern English from the PPCHE corpus (Kroch, 2020) and the PCMEP corpus (Zimmermann, 2015), as seen in Figure 1, bottom. The connection to WF enables a model comparison capable of discriminating between the effects of multiple evolutionary forces. Results show that a model with error in understanding and in production provides a better fit than the null hypothesis of pure drift ($p < 0.001$) and models with selection ($\mathrm{BIC}_{\mathrm{sel}} - \mathrm{BIC}_{\mathrm{err}} \gg 10$).

In summary, we find that this model provides both a framework for grammar change through cultural transmission that reproduces features like the directionality of change and a robust quantitative method for testing causal hypotheses in historical change. Taken together, we gain insights into the components of cultural evolution that were responsible for real-world diachronic phenomena.

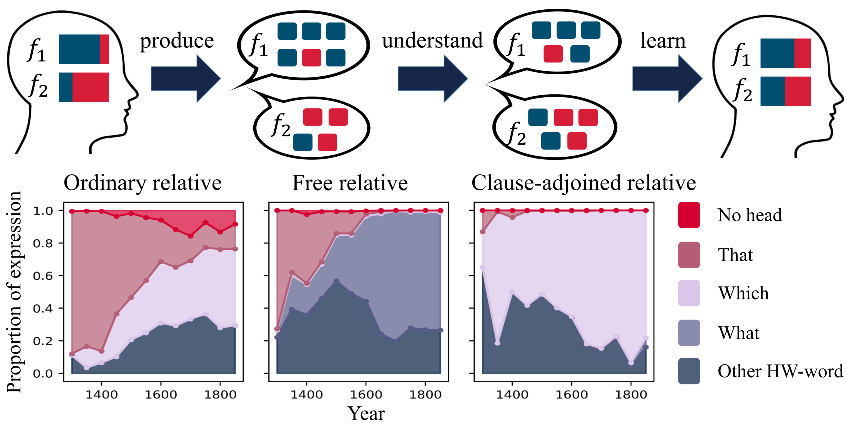


Figure 1. **Top:** Schematic representation of the model for a language with two functions ($F = 2$, f_1 and f_2) and two expressions ($E = 2$, red and blue). A speaker produces utterances for both functions using either expression with a probability dictated by their grammar. The learner then understands those utterances imperfectly, which may lead to some of them being assigned to the wrong function. They then use the utterances to infer the grammar. **Bottom:** Time series of the usage of five different phrases ($E = 5$) as heads of three types relative clauses ($F = 3$) in PPCHE and PCMEP between the years 1300 and 1850. Our model shows that mutation due to error in production (with a yearly rate of $\epsilon = 0.0009$) and in understanding (with a yearly rate of $\eta = 0.00015$) explain the behaviour of the data better than pure drift ($p < 0.001$) or selection ($\mathrm{BIC}_{\mathrm{sel}} - \mathrm{BIC}_{\mathrm{err}} \gg 10$).

References

Griffiths, T., & Kalish, M. (2007). Language evolution by iterated learning with bayesian agents. *Cognitive Science*, *31*, 441–480.

Haspelmath, M. (1999). Why is grammaticalization irreversible?

Kirby, S., Dowman, M., & Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences*, *104*(12), 5241-5245.

Kroch, A. (2020). *Penn Parsed Corpora of Historical English LDC2020T16.* Philadelphia: Linguistic Data Consortium.

Newberry, M., Ahern, C., Clark, R., & Plotkin, J. (2017). Detecting evolutionary forces in language change. *Nature*, *551*, 223-226.

Reali, F., & Griffiths, T. L. (2010). Words as alleles: connecting language evolution with bayesian learners to models of genetic drift. *Proc. R. Soc. B*, *277*, 429-–436.

Smith, K. (2009). Iterated learning in populations of Bayesian agents. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 31).

Zimmermann, R. (2015). *The Parsed Corpus of Middle English Poetry (PCMEP).* University of Manchester, Department of Linguistics and English Language.

Simulating Representational Communication in Vervet Monkeys using Agent-Based Simulation

Ankit Gupta*[1,2], Francis Steen[3]

*Corresponding Author: guptaa23@msu.edu
[1]Department of Computer Science and Engineering & [2]Program in Ecology, Evolution, and Behavior, Michigan State University, East Lansing, MI, United States;
[3]Department of Communication, UCLA, Los Angeles, CA, United States

1. Background and Motivation

Human languages are composed in part by discrete patterns of sound that can be linked to meanings through social learning. What are the evolutionary pressures that drove the development of referential vocalization? In this project, we explore the survival costs and benefits of referential alarm calls in a savanna-dwelling primate. We base our model on the discovery that the African vervet monkey (Chlorocebus pygerythrus) gives acoustically different alarm calls to different predators (Struhsaker, 1967; Seyfarth & Cheney, 1980a, b, c), conditional on context (Deshpande et al., 2023). In the following, we simulate the costs that vervets incur by monitoring their environment for predators, issuing alarm calls, running to escape potential predators, and foregoing foraging in favor of seeking refuge, and track the survival outcomes. Our goal is to determine the envelope of evolvability of representational signaling in the parameter space of a troop of primates in an African savannah environment.

2. Simulation Logic and Methodology

Initiating with a population of vervets and their respective predators, the simulation is guided by the dual parameters of hunger and fear level amongst the vervets. Their quest to diminish hunger through foraging is complicated by the appearance of predators, which elicits a fear response. This encounter prompts vervets to either persist in their foraging efforts or retreat to safe zones (such as stony ground, trees, or bushes), alongside broadcasting an alarm call with a limited radius of efficacy. Predators, meanwhile, meander through the ecosystem, seeking out vervets to maintain their energy through predation. Data was collected across three alarm potency levels: 0 (no call), 1 (general alarm, directing vervets to the nearest refuge - not necessarily safe), and 2 (specific alarm, guiding vervets to the appropriate refuge), alongside variations in vervet resource levels to assess the alarm call's enduring impact.

3. Current Results & Conclusion

Analyzing simulation data from 30 replicates across 50 generations, an interesting pattern emerged. Despite frequent extinctions, vervets with zero alarm potency were observed to survive through significantly more generations compared to those with higher potency levels (Figure 1), as evidenced by the results of the Kruskal-Wallis test, ($H(2) = 77.83$, $p < .001$). Investigating the root causes behind early extinctions by examining average energy levels and predation rates across alarm potency levels for each resource condition also revealed a significant effect of alarm calls. Specifically, vervets without alarm calls had higher energy levels ($H(2) = 196.26$, $p < .001$) and lower predation rates ($H(2) = 10.49$, $p < .01$) compared to those with more potent alarm calls.

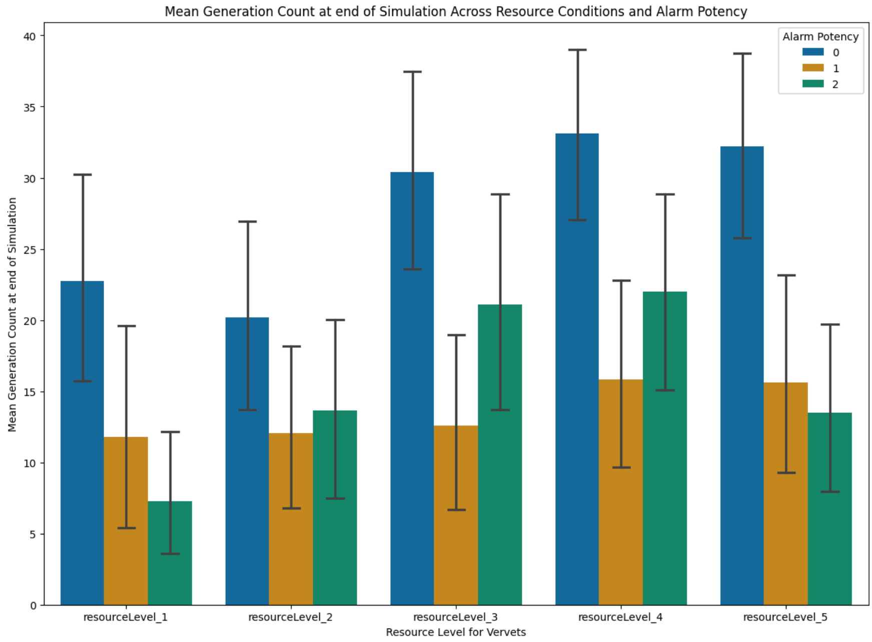


Figure 1: Plot comparing average generation count vervets survive across alarm potency levels across resource conditions. Error bars represent 95% confidence interval.

The aforementioned resource experiment findings on alarm calls prompt further study into variables like predation success probability, energy decay rates, and scanning frequency to reach more conclusive results about alarm call efficacy.

Acknowledgements

This research received support from the Google Summer of Code as part of an open-source project at Red Hen Lab in 2021 and 2022. We also thank Prof. Emily Dolson at Michigan State University for her valuable inputs for the data collection and data analysis.

References

Deshpande, A., van de Waal, E. & Zuberbühler, K. Context-dependent alarm responses in wild vervet monkeys. Anim Cogn 26, 1199–1208 (2023).

Seyfarth RM, Cheney DL, Marler P (1980a) Vervet monkey (Cerco- pithecus aethiops) alarm calls: semantic communication in a free ranging primate. Anim Behav 28:1070–1094

Seyfarth RM, Cheney DL, Marler P (1980b) Vervet monkey alarm calls: semantic communication in a free-ranging primate. Anim Behav 28:1070–1094.

Seyfarth RM, Cheney DL, Marler P (1980c) Monkey responses to three different alarm calls: evidence of predator classification and semantic communication. Science 210:801–803.

Struhsaker TT (1967) Auditory communication among vervet mon- keys (Cercopithecus aethiops). In: Social communication among primates, pp 281–324

Social Ambiguity and Multimodal Interactions in Guinea baboons

Lise Habib-Dassetto[*1,2,3], Mathilde Adet[4], Jules Cauzinille[1,2,5], Arthur Marques[3], Alban Lemasson[6,7], Cristel Portes[3], Julie Gullstrand[1], And Marie Montant[1]

[*]Corresponding Author: habiblise@gmail.com
[1] Aix Marseille Univ, CNRS, CRPN-UMR 7077, Marseille, France
[2] Aix Marseille Univ, CNRS, ILCB
[3] Aix Marseille Univ, CNRS, LPL, Aix-en-Provence, France
[4] Le Mans Univ, France
[5] Aix Marseille Univ, CNRS, LIS-UMR 7020, Marseille, France
[6] Rennes Univ, Normandie Univ, CNRS, EthoS (Ethologie animale et humaine)-UMR 6552, Paimpont, France
[7] Institut Universitaire de France

Many linguistics, comparative psychologists and ethologists consider human language to result mainly from our general communicative abilities and complex, multimodal, structured, and flexible interactional system (i.e., a coherent, integrated set of behaviors occurring between two or more individuals), some of whose characteristics may be shared by other primate species (Beckner et al., 2009; Heesen & Fröhlich, 2022; Levinson, 2019). Therefore, the study of complex interactional abilities (e.g., joint action, turn-taking and/or repair mechanisms) in non-human primates (NHP) should help us to understand the evolution of communication (Heesen & Fröhlich, 2022). According to the "social complexity hypothesis for communicative complexity" (Freeberg et al., 2012), animals who live in complex social environments develop complex communication systems. Indeed, social life (e.g., size of the group, reproductive system, hierarchical system) acts as a selection pressure for the evolution of vocal signals' configuration (for reviews: Cheney & Seyfarth, 2018; Lemasson et al., 2022). Rebout et al. (2020) showed in macaques that compared to intolerant species, tolerant species, who experience more uncertain social interactions, display greater vocal diversity and flexibility. To our knowledge, there are no studies investigating the link between intra-specific features of sociality (e.g., social status, affiliative relationships, hierarchy) and structures of interactions.

Our study aims at investigating the intra-specific effect of sociality on the complexity of multimodal communicative interactions in dyads of Guinea baboons. This species lives in multi-males/multi-females and multi-levels social groups: a group is the sum of several harems, and males, who do not disperse, tend to be tolerant with the other males and sometimes share a harem (Dal Pesco & Fischer, 2018, 2020). Hierarchy is not strictly linear, but males dominate females. We preliminarily tested two social parameters that seem important in that species: the stability of the relationship (determined by the frequency of interactions) and the sex(es) involved in the interaction, as the stakes differ between dyads of males (MM), females (FF), and both sexes (FM). Drawing on Rebout et al. (2020), we consider that these two social parameters will determine whether the outcome of an interaction is more or less uncertain. We hypothesize that this uncertainty should influence the structure of the interaction.

Using focal sampling method (Altmann, 1974), we filmed 66h of a group of 18 Guinea baboons housed at the Primatology Station of Rousset-sur-Arc (France, CNRS). Within a repertoire of 81 multimodal units (vocalizations, gestures, facial expressions, other non-vocal behaviors), we coded the components units of 370 sequences of dyadic interactions on the software BORIS (Friard & Gamba, 2016). A "communicative interaction" was composed of at least one directional signal (Liebal et al., 2004; Pollick & de Waal, 2007). To address the "social complexity hypothesis", we analyzed sequences structure considering their length (number of units), diversity (of units) and their temporal organization (number of units per second and inter-individual overlaps).

Out results show that inter-sex interactions (FM), which always present a sexual stake and are therefore less uncertain than intra-sex interactions (FF and MM), are composed of the weaker diversity of units. We also show that MM interactions have more units per second and more inter-individual overlaps than the other types of dyadic interactions (FF and FM). In the same way, the less frequently individuals interact with each other, the faster they interact and the more their units overlap. In addition to the uncertainty of the relationship, it might be that MM interactions are influenced by the risk interactants take. Relationships between males can indeed lead to dangerous agonistic behaviors, because of potential hierarchical stakes. Three hypotheses could explain the acceleration and overlap in the interactions between males and/or between individuals interacting rarely: (1) the increase in risk would induce an increase in stress and therefore would speed-up the interaction (Lemasson et al., 2010), (2) individuals interacting rarely (especially males) would choose not to leave space for a potential response from the other interactant (Henry et al., 2015; Katsu et al., 2019 on how the quality of a relationship can influence inter-unit delays), (3) as suggested by Pougnault et al. (2022) individuals might use overlap instead of turn-taking as a demonstrative strategy. These preliminary results require further analyses, especially on an individual scale, before we can draw conclusions about their link with the "social complexity hypothesis".

Acknowledgements

This work is supported by the Institute of Convergence ILCB (ANR-16-CONV-0002).

References

Altmann, J. (1974). Observational Study of Behavior: Sampling Methods. *Behaviour*, *49*(3–4), 227–266. https://doi.org/10.1163/156853974X00534

Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., Holland, J., Ke, J., Larsen-Freeman, D., & Schoenemann, T. (2009). Language Is a Complex Adaptive System: Position Paper. *Language Learning*, *59*(s1), 1–26. https://doi.org/10.1111/j.1467-9922.2009.00533.x

Cheney, D. L., & Seyfarth, R. M. (2018). Flexible usage and social function in primate vocalizations. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(9), 1974–1979. https://doi.org/10.1073/pnas.1717572115

Dal Pesco, F., & Fischer, J. (2018). Greetings in male Guinea baboons and the function of rituals in complex social groups. *Journal of Human Evolution*, *125*, 87–98. https://doi.org/10.1016/j.jhevol.2018.10.007

Dal Pesco, F., & Fischer, J. (2020). On the evolution of baboon greeting rituals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1805), 20190420. https://doi.org/10.1098/rstb.2019.0420

Freeberg, T. M., Dunbar, R. I. M., & Ord, T. J. (2012). Social complexity as a proximate and ultimate factor in communicative complexity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1597), 1785–1801. https://doi.org/10.1098/rstb.2011.0213

Friard, O., & 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. https://doi.org/10.1111/2041-210X.12584

Heesen, R., & Fröhlich, M. (2022). Revisiting the human ‘interaction engine’: Comparative approaches to social action coordination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1859), 20210092. https://doi.org/10.1098/rstb.2021.0092

Henry, L., Craig, A. J. F. K., Lemasson, A., & Hausberger, M. (2015). Social coordination in animal vocal interactions. Is there any evidence of turn-taking? The starling as an animal model. *Frontiers in Psychology*, *6*. https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2015.01416

Katsu, N., Yamada, K., Okanoya, K., & Nakamichi, M. (2019). Temporal adjustment of short calls according to a partner during vocal turn-taking in Japanese macaques. *Current Zoology*, *65*(1), 99–105. https://doi.org/10.1093/cz/zoy077

Lemasson, A., Coye, C., Rebout, N., & Levréro, F. (2022). The evolution of primate vocal communication: A social route. *Revue de Primatologie*, *13*, Article 13. https://doi.org/10.4000/primatologie.14835

Lemasson, A., Ouattara, K., Bouchet, H., & Zuberbühler, K. (2010). Speed of call delivery is related to context and caller identity in Campbell's monkey males. *Naturwissenschaften*, *97*(11), 1023–1027. https://doi.org/10.1007/s00114-010-0715-6

Levinson, S. C. (2019). Interactional Foundations of Language: The Interaction Engine Hypothesis. In P. Hagoort (Ed.), *Human Language* (pp. 189–200). The MIT Press. https://doi.org/10.7551/mitpress/10841.003.0018

Liebal, K., Call, J., & Tomasello, M. (2004). Use of gesture sequences in chimpanzees. *American Journal of Primatology*, *64*(4), 377–396. https://doi.org/10.1002/ajp.20087

Pollick, A. S., & de Waal, F. B. M. (2007). Ape gestures and language evolution. *Proceedings of the National Academy of Sciences*, *104*(19), 8184–8189. https://doi.org/10.1073/pnas.0702624104

Pougnault, L., Levréro, F., Leroux, M., Paulet, J., Bombani, P., Dentressangle, F., Deruti, L., Mulot, B., & Lemasson, A. (2022). Social pressure drives "conversational rules" in great apes. *Biological Reviews*, *97*(2), 749–765. https://doi.org/10.1111/brv.12821

Rebout, N., De Marco, A., Lone, J.-C., Sanna, A., Cozzolino, R., Micheletta, J., Sterck, E. H. M., Langermans, J. A. M., Lemasson, A., & Thierry, B. (2020). Tolerant and intolerant macaques show different levels of structural complexity in their vocal communication. *Proceedings of the Royal Society B: Biological Sciences*, *287*(1928), 20200439. https://doi.org/10.1098/rspb.2020.0439

Systemic structure of kinship is shaped by evolutionary processes

Maisy Hallam*[1], Fiona Jordan[2], Simon Kirby[1], and Kenny Smith[1]

*Corresponding Author: maisy.hallam@ed.ac.uk
[1]Centre for Language Evolution, University of Edinburgh, Edinburgh, UK
[2]Department of Anthropology and Archaeology, University of Bristol, Bristol, UK

Kinship terminology is a category system that groups and distinguishes relatives. The number of terms and which relatives are categorised together varies cross-linguistically. For instance, the English kin term *uncle* groups parents' brothers, but the same relatives are split into three categories in Hindi: *cācā* 'father's younger brother', *tāū* 'father's older brother', and *māmā* 'mother's brother'. However, this variation is constrained (Murdock, 1970). Similar categories are distinguished in unrelated languages, and not all theoretically possible categories are attested (Nerlove and Romney, 1969).

What underlies these constraints on diversity? Category systems have been proposed to maximise communicative efficiency (Kemp et al., 2018). Kemp and Regier (2012) show that kinship systems in natural languages near-optimally balance simplicity with informativeness, meaning they tend to have the simplest possible grammar given the number and specificity of the kin terms in the language. Further to this, Passmore et al. (2021) suggest that emergent kin categories are constrained by internal co-selection: an evolutionary process where terminological changes in one generation of the kinship paradigm co-occur with parallel changes in other generations, increasing system-wide predictability. For instance, the collapse of a distinction in Ego's parents' generation may lead to a related collapse in Ego's generation – in Latin and Italian, the merging of terms for mother and father's brothers (*patruus* and *avunculus* collapse to *zio*) was accompanied by a parallel merger in the terms for their children (*frater patruelis*

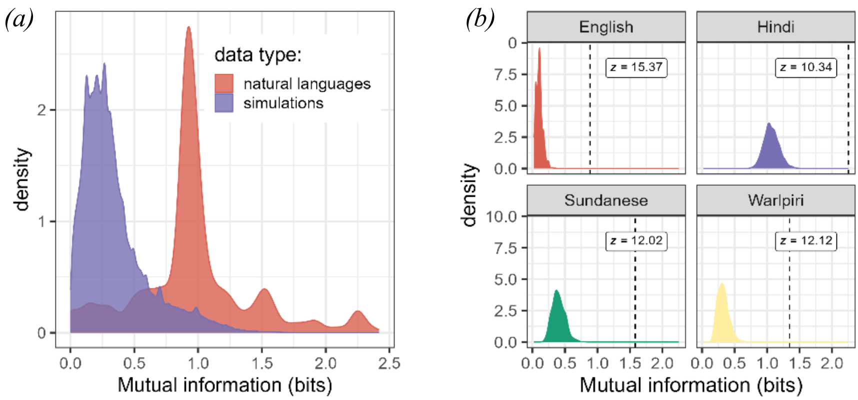


Figure 1. (a) Distribution of mutual information across all natural languages and all simulations. Mutual information is substantially lower in the simulated dataset. **(b)** Distribution of mutual information across simulations for a sample of languages. Dashed line marks the mutual information of the attested kinship system; *z*-score is given relative to the mean of the null distribution (i.e. the distribution of each system's simulated counterparts). For all four languages, mutual information of the natural language is greater than we would expect to arise by chance ($z > 1.96$).

and *filius consobrinus* collapse to *cugino*). Here, we investigate whether kinship systems truly exhibit this predictive structure between generations of kin.

We measured predictive structure between relatives in Ego's generation (i.e. terms for one's siblings and cousins) and Ego's parents' generation (i.e. terms for one's parents and their siblings) as mutual information: an information theoretic measure of how much we can know about the terms in one generation by observing the other. Using kinship terminology data from Kinbank (Passmore et al., 2023) for a sample of 544 languages, we tested whether kinship systems have higher mutual information than chance via a permutation analysis. The mutual information of each language's kinship system was compared to simulated baselines that randomly redistributed kin terms within the paradigm, maintaining the number of terms in each generation but scrambling any predictive structure.

We found 458 kinship systems (84%) had significantly greater mutual information between generations than would be expected if kin terms were distributed randomly ($z = 2.34$, $p < 0.05$) (Figure 1a). Looking at individual kinship systems, we found mutual information is substantially greater than their simulated counterparts (Figure 1b). This tendency to structure kin terms in a predictable way suggests a selective pressure for internal co-selection.

We propose that the internal co-selection process is adaptive because it facilitates the trade-off between simplicity and informativeness (Kemp and Regier, 2012): the more structural information we can predict, the more cognitive resources can be invested in finer-grain kin category distinctions.

Acknowledgements

The first author gratefully acknowledges funding from the Economic and Social Research Council [grant number ES/P000681/1].

References

Kemp, C., & Regier, T. (2012). Kinship Categories Across Languages Reflect General Communicative Principles. *Science*, *336*(6084), 1049–1054.

Kemp, C., Xu, Y., & Regier, T. (2018). Semantic Typology and Efficient Communication. *Annual Review of Linguistics*, *4*(1), 109–128.

Murdock, G. P. (1970, April). Kin Term Patterns and Their Distribution. *Ethnology*, 9, 165.

Nerlove, S. and Romney, A.K. (1967), Sibling Terminology and Cross-Sex Behavior. *American Anthropologist*, 69: 179-187.

Passmore, S., Barth, W., Quinn, K., Greenhill, S. J., Evans, N., & Jordan, F. M. (2021). Kin Against Kin: Internal Co-selection and the Coherence of Kinship Typologies. *Biological Theory*, *16*(3), 176–193.

Passmore, S., Barth, W., Greenhill, S. J., Quinn, K., Sheard, C., Argyriou, P., Birchall, J., Bowern, C., Calladine, J., Deb, A., Diederen, A., Metsäranta, N. P., Araujo, L. H., Schembri, R., Hickey-Hall, J., Honkola, T., Mitchell, A., Poole, L., Rácz, P. M., Jordan, F. M. (2023). Kinbank: A global database of kinship terminology. *PLOS ONE*, *18*(5), e0283218.

Delineating the field of language evolution research: A quantitative study of submission types and peer-review patterns at the Joint Conference on Language Evolution (JCoLE)

Stefan Hartmann*[1], Yannick Jadoul[2], Rie Asano[3], Evelina Daniela Rodrigues[4], Daria Valente[5], Andrea Ravignani[2], and Slawomir Wacewicz[6]

*Corresponding Author: hartmast@hhu.de
[1]German Department, Heinrich-Heine-Universität Düsseldorf, Düsseldorf, Germany
[2]Department of Human Neurosciences, Sapienza University of Rome, Rome, Italy
[3]Advanced Comprehensive Research Organization, Teikyo University, Tokyo, Japan
[4]William James Center for Research, ISPA–Instituto Universitário, Lisbon, Portugal
[5]Department of Life Sciences and Systems Biology, University of Turin
[6]Center for Language Evolution Studies (CLES), Nicolaus Copernicus University, Torun, Poland

Like any field of research (Kuhn, 1970), language evolution research is in constant flux. As such, it seems promising to find bottom-up answers on what is seen as central, and what as more peripheral, in current language evolution research. In this paper, we draw on submission data from the Joint Conference on Language Evolution (JCoLE 2022) to address this question in more detail. Held in September 2022, JCoLE was a joint enterprise of the field's three main conferences (Evolang, Protolang, and Evolinguistics). Following up on previous scientometric studies on Evolang (Bergmann & Dale, 2016, Wacewicz et al., 2023), we use aggregated, anonymized data of the submissions and the peer-review results.

We grouped all submissions by the main research field, as indicated by the submitting author. The core research areas mentioned on the Evolang website are not represented equally. Perhaps unsurprisingly, more than a third of all submissions (37.5%) fall into the linguistics category. Cognitive science is the second most chosen topic (16.9%). Modelling, psychology, primatology, neuroscience, biology follow in this order. When it comes to the relevance scores given by reviewers, the data does not show strong differences between fields. Nonetheless some trends emerge: there is higher consensus in classifying research from biology as relevant; other fields, by contrast, show a broad range of relevance ratings. As could be expected, this higher dispersion is especially true for the many submissions classified under linguistics and cognitive science, but also for e.g. primatology. Papers from psychology and neuroscience show a comparatively lower mean relevance score.

Working with 150-word, Porter-stemmed summaries of submissions (Porter, 1980; implemented in the NTLK Python library, Bird, Klein, & Loper, 2009), we calculated their *tf-idf* score vectors (term frequency-inverse document frequency). We calculated the pairwise cosine similarity between documents, and used these similarity measures as the edge weights in an undirected graph. On this graph, we ran an off-the-shelf community detection algorithm (Clauset-Newman-Moore greedy modularity maximisation, implemented in the NetworkX Python library; Hagberg, Schult, & Swart, 2008) to find 15 disjunct groups of submissions, clustered by the terms used in these author-provided summaries (Figure 1). To get an idea of the topics of submissions within each group, we removed stopwords from the summaries and again used the tf-idf to determine the most relevant words in each group of submissions. The main finding is that differences between groups in the assessed relevance of the submissions are small, testifying to the inherent multidisciplinary nature of language evolution as a research field. The fact that no research topic is entirely dominated by submissions from a single main research field further highlights this point.

Finally, our investigation of the evaluation of paper types (i.e., empirical, modelling, and theoretical work) shows that the community strongly values empirical work: the rejection rate was about 10% for empirical and modelling papers and about 24% for theoretical papers. This indicates that language evolution is considered an empirically tractable phenomenon by the community members. There is also a substantial and growing tendency to re-use existing data, with roughly 25% of submissions reporting studies working with data already available from databases and corpora. As for experimental work, empirical research using communication game setups is considered highly relevant despite the theoretical concerns with using modern humans as participants. Overall, then, our results offer a simple snapshot of language evolution as a dynamically developing field. As new questions, theories and methods keep emerging, the scope of language evolution research and the concept of language evolution will keep adapting.

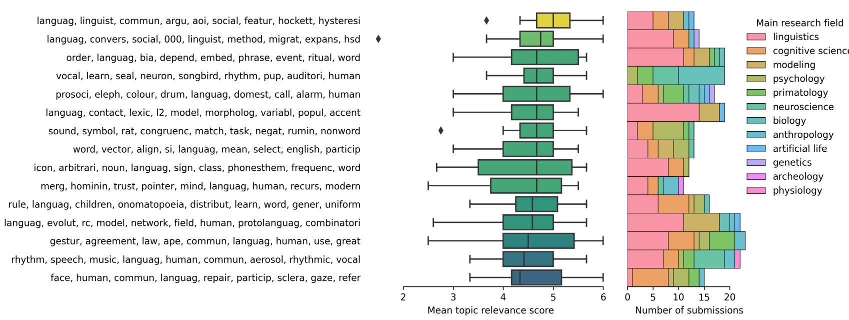


Figure 1. Groups of submissions detected based on the 10 most relevant terms. The boxplots (left) show the distribution of mean relevance scores within each group, and the barplots (right) show the composition of the "main research field" of the group's submissions

Funding

SW was supported by the Polish National Science Center under grant agreement UMO-2019/34/E/HS2/00248. RA was supported by Mext 4093 (Evolinguistics), grant numbers JP22H04897, JP17H06378-383. EDR was supported by the Portuguese national funding agency for science, research, and technology (SFRH/BD/138406/2018).

References

Bergmann, T., & Dale, R. (2016). A sociometric analysis of Evolang: Intersections and authorships. In S. G. Roberts, C. Cuskley, L. McCrohon, L. Barceló-Coblijn, O. Fehér, & T. Verhoef (Eds.), *The Evolution of Language. Proceedings of the 11th International Conference* (pp. 79–86).

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit.* O'Reilly Media, Inc.

Hagberg, A. A., Schult, D. A., & Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. In G. Varoquaux, T. Vaught, & J. Millman (Eds.), *Proceedings of the 7th Python in Science conference* (p. 11-15). Pasadena, CA USA.

Kuhn, T. S. (1970). *The Structure of Scientific Revolutions* (2 ed.). Chicago: The University of Chicago Press.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130–137.

Wacewicz, S., Sibierska, M., Placiński, M., Szczepańska, A., Poniewierska, A., Ng, Y., & Żywiczyński, P. (2023). The scientometric landscape of Evolang: A comprehensive database of the Evolang conference. *Journal of Language Evolution*, lzad003.

Morphological analysis of vocal communication in *Homo naledi*

John Hawks*[1,2]

*Corresponding Author: jhawks@wisc.edu
[1]Department of Anthropology, University of Wisconsin-Madison, USA
[2]Centre for the Exploration of the Deep Human Journey, University of the Witwatersrand, South Africa

Most considerations of the fossil record relevant to human language have relied upon a simplified view of human origins. Many reviews have considered *Australopithecus*, *Homo erectus*, Neanderthals, and modern humans as stages, arguing that cranial anatomy and material culture provide evidence of symbolic language only in modern human contexts (Hauser et al., 2014). Others have emphasized that Neanderthals had adaptations consistent with spoken language (Johanson, 2015).

Discoveries of the last two decades have markedly increased the level of data and diversity of species known in the hominin record. Diverse species of *Homo*, including *Homo floresiensis* (Brown et al., 2004) and *Homo naledi* (Berger et al., 2015), persisted into the later Pleistocene with small brain size, but with endocast form similar to modern humans in some important ways. Here I focus on the anatomy of *Homo naledi* relevant to vocal communication and language. The current fossil record of this species now numbers more than 2000 fossil fragments, representing a minimum of 25 individuals from four localities within the Rising Star cave system of South Africa (Berger et al., 2015; Hawks et al., 2017; Brophy et al., 2021; Berger et al., 2023). The remains from the Dinaledi Chamber have been placed between 335,000 and 241,000 years ago (Dirks et al., 2017). The context of the fossil material suggests that *H. naledi* was making repeated use of the dark zone of this cave system, which may have entailed some cultural tradition or transmission of knowledge across individuals and generations (Berger et al., 2023).

The endocast volume of *Homo naledi* crania ranges from 450 ml to 610 ml, overlapping with *Australopithecus* and early *Homo* species. Endocasts of three *H. naledi* individuals preserve evidence from the left prefrontal cortex relevant to the morphology of Broca's area. They show that *H. naledi* had a configuration similar to modern humans and other recent fossil *Homo* (Holloway et al., 2018), although different from some early *H. erectus* fossil endocasts (Ponce de Léon et al., 2021). No hyoid bone fragments have yet been identified, but fossil ear incudes are known and are similar in some ways to the morphology to the *Paranthropus* incus (Elliott et al., 2018). Some past work has considered the basicranial form to be relevant to vocalization ability, and reconstruction of the basicranium of *H. naledi* shows a morphology that is broadly similar to early *H. erectus* with a somewhat greater degree of basicranial flexion. The auditory canal of *H. naledi* diverges from those of *Australopithecus* or *Paranthropus* in size and shape, with greater similarity to other Middle Pleistocene *Homo* in shape and a smaller size that may reflect scaling with skull size.

Fossil evidence has many weaknesses as a test for vocal behavior, and the most important anatomical correlates of language in the brain do not leave a fossil trace. However, *H. naledi* presents an anatomical picture that is nearly as complete as that known for Neanderthals, and more complete than *H. erectus*, in a Middle Pleistocene context when early *H. sapiens* was also extant. The anatomy suggests that this species had some features that are associated with vocal communication in recent humans and are not found in early *H. erectus*. These features may reflect either homology or convergence depending on the phylogenetic hypothesis used.

Acknowledgements

Funding for this project is acknowledged from the National Geographic Society, John Templeton Foundation, Fulbright U.S. Scholar Program, and the University of Wisconsin-Madison. The Lee R. Berger Foundation is acknowledged for access to the site and facilities, and the University of the Witwatersrand for curation of the fossil material.

References

Berger, L. R., Hawks, J., de Ruiter, D. J., Churchill, S. E., Schmid, P., et al. (2015). *Homo naledi*, a new species of the genus *Homo* from the Dinaledi Chamber, South Africa. *eLife*, *4*, e09560.

Berger, L. R., Makhubela, T., Molopyane, K., Krüger, A., Randolph-Quinney, P., et al. (2023). Evidence for deliberate burial of the dead by Homo naledi. *eLife*, *12*.

Brophy, J. K., Elliott, M. C., Ruiter, D. J. D., Bolter, D. R., Churchill, S. E., et al. (2021). Immature Hominin Craniodental Remains From a New Locality in the Rising Star Cave System, South Africa. *PaleoAnthropology*.

Brown, P., Sutikna, T., Morwood, M. J., Soejono, R. P., Jatmiko, Wayhu Saptomo, E., & Awe Due, R. (2004). A new small-bodied hominin from the Late Pleistocene of Flores, Indonesia. *Nature*, *431*(7012), Article 7012.

Dirks, P. H., Roberts, E. M., Hilbert-Wolf, H., Kramers, J. D., Hawks, J., et al. (2017). The age of *Homo naledi* and associated sediments in the Rising Star Cave, South Africa. *eLife*, *6*, e24231.

Elliott, M. C., Quam, R., Nalla, S., de Ruiter, D. J., Hawks, J., & Berger, L. R. (2018). Description and analysis of three Homo naledi incudes from the Dinaledi Chamber, Rising Star cave (South Africa). *Journal of Human Evolution*, *122*, 146–155.

Hauser, M. D., Yang, C., Berwick, R. C., Tattersall, I., Ryan, M. J., et al. (2014). The mystery of language evolution. *Frontiers in Psychology*, *5*.

Hawks, J., Elliott, M., Schmid, P., Churchill, S. E., de Ruiter, D. J., Roberts, E. M., et al. (2017). New fossil remains of *Homo naledi* from the Lesedi Chamber, South Africa. *eLife*, *6*, e24232.

Holloway, R. L., Hurst, S. D., Garvin, H. M., Schoenemann, P. T., Vanti, et al. (2018). Endocast morphology of *Homo naledi* from the Dinaledi Chamber, South Africa. *Proceedings of the National Academy of Sciences*, *115*(22), 5738–5743.

Johansson,S. (2015). Language abilities in Neanderthals. *Annual Review of Linguistics*, 1(1), 311-332.

Ponce de León, M. S., Bienvenu, T., Marom, A., Engel, S., Tafforeau, P., et al. (2021). The primitive brain of early *Homo*. *Science*, *372*(6538), 165–171.

Simplicity and informativeness in the evolution of combinatorial structure

Matthias Hofer[*1], Simon Kirby[2], and Roger Levy[1]

[1]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, USA
[2]Centre for Language Evolution, University of Edinburgh, UK
[*]correspondence to mhofer@mit.edu

Symbol systems—including language, music and pictorial diagrams—are crucial for the storage and transmission of human knowledge, ultimately underpinning our capacity for cumulative culture. A central feature underwriting these systems' success is combinatorial structure: the reuse of building blocks to compose new concepts or ideas. Despite its apparent advantages, not all symbol systems rely on combinatorial structure, as evidenced for example by work emerging sign languages (Sandler, Meir, Padden, & Aronoff, 2005; Sandler, Aronoff, Meir, & Padden, 2011), indicating that combinatorial structures may only develop under certain conditions, rather than being an inherent property of symbol systems.

A body of recent work in evolutionary linguistics proposes that features like compositionality or systematicity, developed as efficient solutions to evolutionary trade-offs encountered in language acquisition and usage in cultural transmission (Kirby & Hurford, 2002; Kirby, Tamariz, Cornish, & Smith, 2015; Tamariz & Kirby, 2016; Christiansen & Chater, 2016; Smith, 2021). Such a framework might also account for the emergence of combinatorial structure, whose spontaneous emergence has previously been demonstrated across several domains (Verhoef, Kirby, & de Boer, 2014; Little, Eryılmaz, & de Boer, 2017; Lieck & Rohrmeier, 2021), but the mechanism by which it emerges from learning and use trade-offs remains unexplored.

In this study, we combine iterated learning with a communication game to directly test this *trade-off hypothesis* (Kirby & Tamariz, 2022), while concurrently exploring the role of pressures for communicative efficiency. We hypothesize that the communication task might not only lead to expressivity and help synchronize emerging patterns, but speaker-related pressures in communication may also contribute directly to signal simplification, over and above the learning process.

In our experiment, participants used a digital slide whistle to create signals for a set of visual referents. Two conditions are contrasted: one focused solely on learning and signal reproduction, without communication demands ("learning only"); another condition added a communication game after the learning phase ("learning plus communication"). During the learning phase, participants mem-

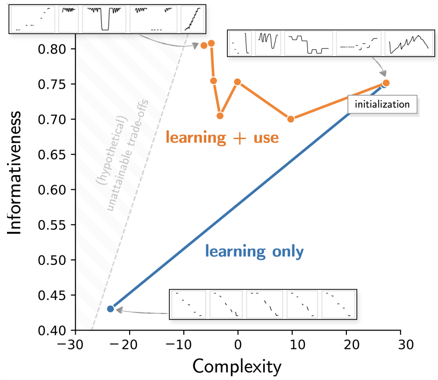


Figure 1. The figure compares the evolution of signaling systems under two conditions. Initially complex and informative (top right), systems evolve differently: without communicative biases (”learning only”), they become overly simple and less expressive. When learning is combined with communication, systems retain informativeness and simplify by developing combinatorial patterns, supporting the *trade-off hypothesis* that combinatoriality arises from balancing complexity and informativeness.

orized and reproduced five distinct whistled signals corresponding to five novel visual referents, which lacked any obvious compositional features that could lead to corresponding structures in the signal space. Signal-referent pairings were also randomized across generations to prevent predictable iconic associations. In the communication phase, participants in the ”learning plus communication” condition were paired to play a reference game where they used their learned signals to refer to selected referents. This process was repeated across 15 chains of participants over five generations, allowing us to observe the evolution of the signaling systems under different conditions.

Consistent with the predictions of the trade-off hypothesis, our findings (see Figure 1) show that combinatorial structure only emerges when languages are subject to pressures from both learning and communication, marking this the first direct test of this theoretical prediction. While signals do become progressively simpler in the learning only condition, they do so to a point of loosing discriminable features while not exhibiting marks of combinatoriality. Additionally, contrary to what a speaker-centered efficiency account might predict, speaker-related simplicity pressures in the communication did not contribute to overall language simplicity during the communication phase.

These findings indicate that combinatorial structures arise from the complex interaction of cognitive biases towards simplicity and informativeness, and align with the cultural evolutionary account of language development. Future research should explore the applicability of these results to different signal modalities and examine the computational basis of these biases through computational modeling.

References

Christiansen, M. H., & Chater, N. (2016). *Creating language: Integrating evolution, acquisition, and processing.* Cambridge, MA: The MIT Press.

Kirby, S., & Hurford, J. R. (2002). The emergence of linguistic structure: An overview of the iterated learning model. In *Simulating the evolution of language* (pp. 121–147). Springer.

Kirby, S., & Tamariz, M. (2022). Cumulative cultural evolution, population structure and the origin of combinatoriality in human language. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1843), 20200319.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102.

Lieck, R., & Rohrmeier, M. (2021). Discretisation and continuity: The emergence of symbols in communication. *Cognition*, *215*, 104787.

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

Sandler, W., Aronoff, M., Meir, I., & Padden, C. (2011). The gradual emergence of phonological form in a new language. *Natural Language & Linguistic Theory*, *29*(2), 503–543.

Sandler, W., Meir, I., Padden, C., & Aronoff, M. (2005). The emergence of grammar: Systematic structure in a new language. *Proceedings of the National Academy of Sciences*, *102*(7), 2661–2665.

Smith, K. (2021). *How language learning and language use create linguistic structure* (Preprint). PsyArXiv.

Tamariz, M., & Kirby, S. (2016). The cultural evolution of language. *Current Opinion in Psychology*, *8*, 37–43.

Verhoef, T., Kirby, S., & de Boer, B. (2014). Emergence of combinatorial structure and economy through iterated learning with continuous acoustic signals. *Journal of Phonetics*, *43*, 57–68.

Using the scale of innovation to study the evolution of language

Annie Holtz*[1]

*Corresponding Author: annie.holtz@ed.ac.uk
[1]Centre for Language Evolution, University of Edinburgh, Edinburgh, United Kingdom

1. Introduction

Researchers who are interested in the cognitive basis of various linguistic phenomena often use experiments to examine how participants interact with linguistic stimuli. In language evolution research, much of this work has focused on how cognitive biases that are present in the minds of individuals come to affect population-wide phenomena, like typological regularities and the pathway of language evolution (Culbertson, 2023).

I review results from previous experiments examining the effects of cognitive biases on linguistic structure across syntax and phonology, and show that most of these cognitive biases can be categorised into two groups, either *category-specific* or *system-wide*. Category-specific biases are motivated by factors external to the linguistic domain in which they apply (e.g. affect syntactic structure but are motivated by semantics/processing), and adherence to these biases can be evaluated on an item-by-item basis (i.e. without reference to the wider linguistic system to which they belong). System-wide biases, on the other hand, are motivated by factors internal to the linguistic domain in which they apply, and adherence to these biases must be evaluated in the context of the wider linguistic system. An example of a category-specific bias is the subject (or agent) first bias (Futrell et al., 2015; Meir et al., 2017), as this is motivated by the accessibility of animate entities (Dahl, 2008; Yamamoto, 1999) and affects the syntactic structure of languages by favouring subject initial word orders (Goldin-Meadow, So, Özyürek, & Mylander, 2008). An example of a system-wide bias is the preference for harmonic (or consistent) order between heads and dependents. This bias is motivated by the compression benefits that harmony affords to a grammar (Culbertson & Kirby, 2016), and causes participants to favour harmonic languages (Culbertson, 2012).

2. Proposal

Crucially, the pattern of evidence for the two types of biases shows that the amount of *innovation* involved in the experimental task is predictive of the emergence of behavioural evidence for the two bias types. Category-specific biases tend to influence participants' behaviour in tasks requiring innovation, whereas system-wide biases are more active in learning-based tasks that require little innovation (Finley & Badecker, 2007; Martin & Peperkamp, 2020; Moreton & Pater, 2012; Motamedi, Wolters, Naegeli, Kirby, & Schouwstra, 2022).

In light of this pattern I introduce the *Scale of Innovation* (see Figure 1) as a conceptual tool for organising experimental contexts and paradigms along a continuum based on the amount of linguistic innovation that is required in each experimental task. On the left are task that rely on learning (low-innovation contexts like memorisation), on the right are tasks that rely on improvisation (high-innovation contexts), and in the middle are tasks that combine learning and improvisation (mixed-innovation contexts like extrapolation). This scale, along with the division of cognitive biases into category-specific and system-wide types, allows researchers to choose appropriate experimental methods to study their chosen cognitive phenomenon.

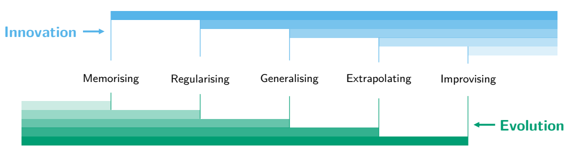


Figure 1. Scale of innovation showing contexts that require more innovation as you move right along the top scale, and the progression of language evolution as you move left along the bottom scale.

To support the order and divisions along the scale I present results from a novel set of artificial language learning experiments that focus on a typological pattern in the noun phrase that exhibits competition between category-specific and system-wide biases (i.e. an instance where the system-wide bias for harmony competes with category-specific biases favouring nonharmonic order). I show that, by manipulating the amount of innovation involved in the experimental task, participants' interaction with this phrase is influenced by either one of the bias types individually (when performing tasks at opposite ends of the scale) or a combination of both (when performing tasks at the centre of the scale).

3. Discussion

In addition to the practical benefits of using the scale during experiment design, there are also potential parallels between the structure of the scale of innovation and different stages of language evolution. This is illustrated in the matched scale at the bottom of Figure 1, where improvisation is matched to the start of language evolution. I discuss these parallels with evidence from language acquisition, studies of young sign languages, and proposed models of language evolution.

Acknowledgements

I would like to thank Jennifer Culbertson and Simon Kirby for useful discussions and feedback during the development of this theory, and subsequent experimental tests. This research was supported by the School of Philosophy, Psychology and Language Sciences full scholarship for a PhD in Linguistics and English Language at the University of Edinburgh.

References

Culbertson, J. (2012). Typological universals as reflections of biased learning: Evidence from artificial language learning. *Language and Linguistics Compass*, *6*(5), 310–329.

Culbertson, J. (2023). Artificial language learning. In J. Sprouse (Ed.), *Oxford handbook of experimental syntax* (pp. 271–300). Oxford University Press.

Culbertson, J., & Kirby, S. (2016). Simplicity and specificity in language: Domain-general biases have domain-specific effects. *Frontiers in Psychology*, *6*, 1964.

Dahl, Ö. (2008). Animacy and egophoricity: Grammar, ontology and phylogeny. *Lingua*, *118*(2), 141–150.

Finley, S., & Badecker, W. (2007). Towards a substantively biased theory of learning. In *Annual meeting of the berkeley linguistics society* (Vol. 33, pp. 142–153).

Futrell, R., Hickey, T., Lee, A., Lim, E., Luchkina, E., & Gibson, E. (2015). Cross-linguistic gestures reflect typological universals: A subject-initial, verb-final bias in speakers of diverse languages. *Cognition*, *136*, 215–221.

Goldin-Meadow, S., So, W. C., Özyürek, A., & 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.

Martin, A., & Peperkamp, S. (2020). Phonetically natural rules benefit from a learning bias: a re-examination of vowel harmony versus disharmony. *Phonology*, *37*(1), 65-90.

Meir, I., Aronoff, M., Börstell, C., Hwang, S.-O., Ilkbasaran, D., Kastner, I., Lepic, R., Ben-Basat, A. L., Padden, C., & Sandler, W. (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.

Moreton, E., & Pater, J. (2012). Structure and substance in artificial-phonology learning, part II: Substance. *Language and Linguistics Compass*, *6*(11), 702–718.

Motamedi, Y., Wolters, L., Naegeli, D., Kirby, S., & Schouwstra, M. (2022). From improvisation to learning: How naturalness and systematicity shape language evolution. *Cognition*, *228*, 105206.

Yamamoto, M. (1999). *Animacy and reference: A cognitive approach to corpus linguistics.* Amsterdam: John Benjamins Publishing Company.

Failures and successes to learn a core conceptual distinction from the statistics of language

Zhimin Hu[1], Jeroen van Paridon[1], and Gary Lupyan[*1]

[*]corresponding author
[1]Department of Psychology, University of Wisconsin-Madison, Madison, USA

Abstract

Generic statements like "tigers are striped" and "cars have radios" communicate information that is, in general, true. However, while the first statement is true *in principle*, the second is true only statistically. People are exquisitely sensitive to this principled-vs-statistical distinction. It has been argued that this ability to distinguish between something being true by virtue of it being a category member versus being true because of mere statistical regularity, is a general property of people's conceptual machinery and cannot itself be learned. We investigate whether the distinction between principled and statistical properties can be learned from language itself. If so, it raises the possibility that language experience can bootstrap core conceptual distinctions and that it is possible to learn sophisticated causal models directly from language. We find that language models are all sensitive to statistical prevalence, but struggle with representing the principled-vs-statistical distinction controlling for prevalence. Until GPT-4, which succeeds.

Keywords: distributional semantics; generics; world models

1. Introduction

People interpret generic statements such as *airplanes have wings*, and *dogs bark* to mean that the named property is, in general, true of the category (Hollander et al., 2009). Other statements of this form, however, such as *airplanes carry passengers* and *dogs wear collars*, while also being judged as generally true, have a decidedly different quality. In a series of papers, Prasada and colleagues (Prasada & Dillingham, 2006; Prasada, 2016; Prasada et al., 2013) drew a distinction between generics that express *principled* properties and generics that express merely *statistical* properties. A statement expressing a principled property such as *airplanes have wings* retains its truthfulness when asked whether it is true because of (or by virtue of) being that thing. For example, in the experiments we describe below, on a scale of -3 = completely false to +3 = completely true, people judged the statement *airplanes have wings* with mean of 2.9. This declines only slightly if asked whether it is true that airplanes have wings *because they are airplanes*

(M=2.6). A statement like *airplanes have passengers* is judged to also be mostly true (M=1.8), but if asked whether airplanes have passengers *because* they are airplanes, the truth estimate drops (M=0.6). Importantly, this key result remains when one controls for confounds such as prevalence and cue-validity, showing that it is not simply an artifact of principled connections being more common or it being harder to come up with counter-examples.

Results like these have been used to argue that people's ability to distinguish between principled and statistical generics requires an *a priori* sensitivity to a distinction between statistical vs. "in-principle" properties. Because there are no structural differences between generics that could inform this distinction, it is thought that the distinction cannot be learned through associations (see Prasada et al., 2013; Haward, Wagner, Carey, & Prasada, 2018), and perhaps cannot even be represented by an associative mechanism (Prasada, 2021).

However, even though generic statements do not encode the principled/statistical distinction in their structure, the distinction might still be captured in the distributional structure of language itself. In this study, we investigated whether the statistical/generic distinction is recoverable from the statistics of language. We did this by predicting human judgments of generic statements from judgments derived from distributional language models. Finding that this distinction can be learned by an associative mechanism from language alone is important for two main reasons. First, it shows that it is *in principle* possible to learn a formal conceptual distinction argued to be unlearnable (and even unrepresentable) by an associative mechanism. Second, it opens the door to asking questions of key interest to the study of language evolution: (1) Are languages structured to facilitate extracting principled item-property relationships, (2) Where in language is such information represented? (3) Are languages not only a *source* of generic information (Rhodes, Leslie, & Tworek, 2012) but do they help structure the very core of our conceptual system?

To anticipate our results, we find that language models are all sensitive to item prevalence. Statements probing frequent item-property combinations like *orange grow on trees* and *kangaroos have pouches* are judged by models as more true than statements probing rarer item-property combinations such as *professors are absent-minded* and *birds are kept in cages*. However, a distinction in truth judgments between principled and statistical relations when controlling for prevalence and cue-validity only appeared for the largest language models we tested.

2. Human ratings

We began by constructing a corpus of 208 generic statements and having them rated on several scales using a procedure adapted from Prasada et al., 2013.

2.1. *Participants*

We recruited 91 native speakers of English residing in the United States through Amazon Mechanical Turk in exchange for a $2 payment. Seven participants were rejected for failing basic attention checks, leaving 84 participants.

2.2. *Procedure*

Participants were asked to judge four different aspects of generic statements, sentences describing properties of objects, people, and animals: (1) *Bare generic truth judgment:* "How true is the following statement: *Airplanes have seatbelts.*"; (2) *By-virtue-of truth judgment*: "How true is the following statement: *Because they are airplanes, airplanes have seatbelts.*"; (3) *Prevalence rating*: "Think of airplanes, how likely are they to have seatbelts?"; (4): *Cue validity rating*: "You learn that [unknown things/people] have wings, how likely is it [are they] to be airplanes?" Each participant was presented with 26 statements of each type. The statements were counterbalanced across participants and rating questions so that no participant rated a given generic more than once.

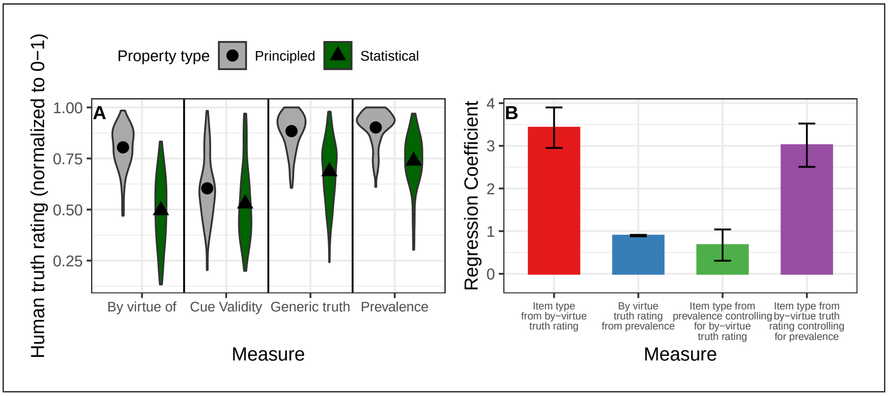


Figure 1. A. Mean human truth ratings for each sentence frame, comparing principled and statistical relationships. B. Regression coefficients (with SEs) showing key relationships between truth ratings, property type, and prevalence (see text.)

2.3. *Results*

Our results, shown in Fig. 1, closely replicate the findings of Prasada et al., 2013. Bare generics ("Airplanes have wings") expressing principled relationships are rated as more true than statements expressing statistical relationships ("Airplanes have passengers") and the same goes for by-virtue-of judgments (but more so). By-virtue truth ratings *are* affected by prevalence (the blue bar in Fig. 1B) and prevalence predicts item-type (principled vs. statistical) when controlling for the by-truth rating (green bar). Importantly, the ability of by-virtue judgments to predict property type (red bar) is nearly undiminished when we control for prevalence (cf. red and purple bars). We will be comparing the model results to this U-shaped pattern of coefficients shown in Fig. 1B.

3. Can the principled/statistical distinction be learned from language itself?

To determine whether distributional models can differentiate between statistical and principled generics, we predicted property type (statistical vs. principled) from the cosine similarity between the target-word and the property.

3.1. *Models*

We tested the language models listed Table 1 using the Huggingface implementations of BERT (Devlin et al., 2018), ALBERT (Lan et al., 2019), DistilBERT (Sanh et al., 2019), RoBERTa (Liu et al., 2019), GPT (Radford et al., 2018), and GPT-2 (Radford et al., 2019). We used the OpenAI APIs for GPT-3.5 and GPT-4.

Table 1. Overview of the models we tested

Model name	Training sources	Size of training corpus	# Number of parameters
BERT (base)	Wiki, books	3.3B tokens (13 GB data)	110M
ALBERT (base-v1)	Wiki, books	3.3B tokens (13 GB data)	11M
Distilbert (base)	Wiki, books	3.3B tokens (13 GB data)	66M
RoBERTa (base)	Wiki, books, web crawl	161 GB data	125M
GPT	Web crawl	800M tokens	110M
GPT-2 (base)	Web crawl, Reddit,	8M documents (40 GB data)	117M
GPT-3.5	Unknown superset of GPT-2	Unknown	Unknown
GPT-4	Unknown superset of GPT-3.5	Unknown	Unknown

3.2. *Methods*

To measure the represented similarity between the target words and their properties, we first needed to obtain their model embeddings. Because the transformer models only generate contextual embeddings, we simulated a decontextualized context by using the ”all but the top” method proposed by (Mu & Viswanath, 2018). This method removes the top k principal components (here, k=7) as computed by sampling additional corpuses of text from the NLI dataset (Bowman et al., 2015) and wiki-103 (Merity et al., 2016). It ensures the resulting embeddings reflect a more contrastive meaning of a given phrase. The models' truth judgment was then operationalized as the cosine similarity between the target-word (e.g., ”airplanes”) and the property (”have wings”).

Because GPT-3.5 and 4 are fine-tuned for question-answering, it was possible to probe their 'knowledge' more directly by having them rate the generics using the same prompt as human participants. The models received the following type prompt: *return only one integer between -3 and 3 where -3 means the sentence is definitely false and 3 means the sentence is definitely true : Because they are airplanes, airplanes have wings*. We tested each of the 208 generics 15 times and averaged the ratings. The variance of this average was less than 0.01.

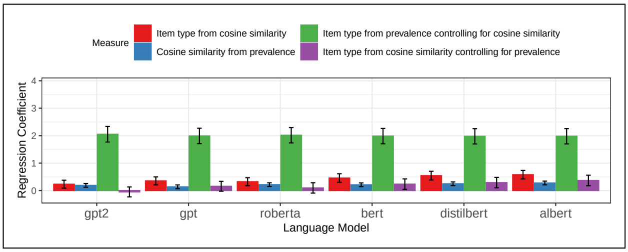


Figure 2. Regression coefficients (with SEs) indicating relationships between item-property cosine-similarity, property-type, and prevalence using the analogous models used in Fig. 2.

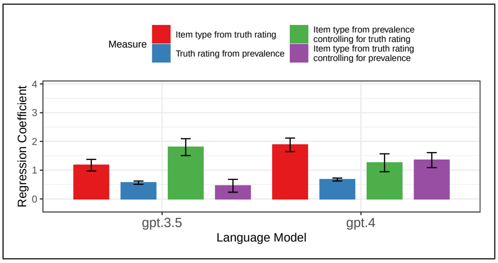


Figure 3. Regression coefficients (with SEs) indicating relationships between model-generated by-virtue-of truth ratings, property-type and human-ratings of prevalence.

4. Results

The basic pattern of results from the cosine similarity analyses is shown in Fig. 2. Across all six models we see the same qualitative pattern. The models distinguish between principled and statistical connections: the similarity between the target word like ‘airplane’ and a principled property like ‘wings’ is greater than a statistical property like ‘passengers (red bars). However, when we control for prevalence, this association largely disappears (purple bars; it is only marginally above 0 in ALBERT). Human truth ratings (especially by-virtue ratings) are much better predictors of property-type than the prevalence ratings. For the models, this is not the case as indicated by the large green bar in comparison to 1B.

Turning to our experiments of GPT-3.5 and GPT-4 in which we were able to directly query their truth judgments, we find a rather different result (Fig. 3).

In GPT-3.5, truth ratings only are barely predictive of property-type when controlling for prevalence (green bar; t=2.05, p=.04), while for GPT-4 they remain strongly predictive, t=5.18, p <.00001). As a complementary analysis, we examined by-item relationships. For each item (e.g., dogs, trampolines, trumpets), we can compare the by-human virtue-of truth judgment for the principled vs. statistical statement, and compare it to the cosine-similarity-based measure for the BERT-type models and to the truth-judgments for the GPT models. We find correlations ranging from .21 for BERT to .28 for DistilBERT. These increase to .49 for GPT-3.5 and to .61 for GPT-4.

5. General Discussion

People know that airplanes have wings and carry passengers, and simultaneously know that the former but not the latter property is part of what *it means* to be an airplane. Since this distinction is not marked in language, it has been thought that it must come from elsewhere such as an innate generative type-token mechanisms (Prasada, 2016). We show here that it is, in principle, possible to learn this distinction from the statistics of language, but it is far from trivial, emerging most clearly only in GPT-4. All tested transformer models trained on English text were sensitive to prevalence as shown by significant associations between prevalence and cosine similarity/model truth judgments. People's judgments too show sensitivity to prevalence which makes sense since it is often a good proxy for whether a relationship is principled or statistical: that *principled* relationships have, on average, considerably higher prevalence than merely *statistical* ones). But human judgments continue to strongly distinguish principled and statistical relationships when prevalence is partialled out–consistent with the view that people base their judgments on causal models, presumably learned from rich multimodal experience (see e.g. Prasada & Dillingham, 2006; Prasada et al., 2013). The failure of language models to distinguish statistical from principled properties once prevalence is partialled out indicates that the models are basing their 'judgments' on statistical co-occurrence. And yet, when we test more recent models such as GPT-3.5 and especially GPT-4, the picture starts to shift consistent with the possibility that lowering next-token prediction error at scale can lead to the models inducing more sophisticated world models (e.g., Li et al., 2022; Mirchandani et al., 2023; Michaelov et al., 2023; Li et al., 2021). Although it is unknown at present what allows GPT-4 to succeed, our experiments provide an in-principle proof that it is possible to induce sophisticated causal models of item-property relations from language alone. Although it is rather unlikely that people learn the distinction between principled and statistical properties from language in the same way, our results hint that input from language may be more instrumental for laying down core conceptual distinctions than previously thought.

References

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 632–642). Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Haward, P., Wagner, L., Carey, S., & Prasada, S. (2018). The development of principled connections and kind representations. *Cognition*, *176*, 255–268.

Hollander, M. A., Gelman, S., & Raman, L. (2009). Generic language and judgements about category membership: Can generics highlight properties as central? *Language and cognitive processes*, *24*(4), 481–505.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. In *International conference on learning representations.*

Li, B. Z., Nye, M., & Andreas, J. (2021). Implicit representations of meaning in neural language models. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)* (pp. 1813–1827).

Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., & Wattenberg, M. (2022). Emergent world representations: Exploring a sequence model trained on a synthetic task. In *The eleventh international conference on learning representations.*

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Merity, S., Xiong, C., Bradbury, J., & Socher, R. (2016). Pointer sentinel mixture models. In *International conference on learning representations.*

Michaelov, J. A., Coulson, S., & Bergen, B. K. (2023). Can peanuts fall in love with distributional semantics? *arXiv preprint arXiv:2301.08731*.

Mirchandani, S., Xia, F., Florence, P., Driess, D., Arenas, M. G., Rao, K., Sadigh, D., Zeng, A., et al.. (2023). Large language models as general pattern machines. In *7th annual conference on robot learning.*

Mu, J., & Viswanath, P. (2018). All-but-the-top: Simple and effective postprocessing for word representations. In *International conference on learning representations.*

Prasada, S. (2016). Mechanisms for thinking about kinds, instances of kinds, and kinds of kinds. In *Core knowledge and conceptual change* (pp. 209–224). New York, NY, US: Oxford University Press.

Prasada, S. (2021). The physical basis of conceptual representation - An addendum to. *Cognition*, *214*, 104751.

Prasada, S., & Dillingham, E. M. (2006). Principled and statistical connections in common sense conception. *Cognition*, *99*(1), 73–112.

Prasada, S., Khemlani, S., Leslie, S.-J., & Glucksberg, S. (2013). Conceptual distinctions amongst generics. *Cognition*, *126*(3), 405–422.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.. (2018). Improving language understanding by generative pre-training. *OpenAI blog*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Rhodes, M., Leslie, S.-J., & Tworek, C. M. (2012). Cultural transmission of social essentialism. *Proceedings of the National Academy of Sciences*, *109*(34), 13526–13531.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Compositionality as one of the enabling factors to communicate conceptualized meaning

Ibuki Iwamura*[1], Hayate Funakura[2], and Takashi Hashimoto[1]

*Corresponding Author: ibuki_iwamura@jaist.ac.jp
[1]Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology, Nomi, Japan
[2]Graduate School of Human and Environmental Studies, Kyoto University, Kyoto, Japan

1. Introduction

Humans linguistically communicate their conceptualized meanings that do not simply refer to the external world (Langacker, 2001, 2002, 2014). People may differently conceptualize the same objective event and reflect the different conceptualizations in linguistic forms. Thus, the meaning of a form is not uniquely determined only from the objective external event. Those who have not acquired how to reflect conceptualization in forms have difficulty in inferring the speaker's subjective conceptualized meanings.

The iterated learning model (ILM), a model for cultural evolution of compositional language simulating intergenerational transmission (Kirby, 2002, among others), does not treat conceptualized meanings since it posits that forms refer to external events. Further, listener agents in ILM receive complete information about both meanings and forms. It is not obvious whether compositionality evolves when conceptualized meanings are communicated as actual linguistic communication. While learners in a holistic language world cannot understand mappings between speaker's conceptualized meanings and holistic forms, compositionality may facilitate learners to infer the mapping.

In this study, we investigate under what condition compositional language is transmitted via a process of cultural evolution when agents can reflect their own conceptualized meaning in forms.

2. Model and Experiment

We simplified the ILM, which is based on definite clause grammar proposed by Kirby (2002), with limited number of letters, 3, and added the conceptualized meaning as agent's internal variable and the ability to reflect it in forms. The conceptualized meaning is represented by a variable added to the semantic representation of predicate-argument structure, $p(\alpha_{1,}\ \alpha_2)/CV$ where p is a predicate, $\alpha_1\ and\ \alpha_2$ are arguments, and CV signifies the conceptualized meaning, which takes only binary values, 0 or 1, for simplicity. Note that this simplification holds the essence of conceptualization since binary options such as active/passive voices and special relationships like "A on B/B below A" are not determined objectively by external events but are decided subjectively through language users' conceptualization. Supposing shared intentionality, learner agents perceive invisible speaker's CV at a certain probability, set as 0.8 in our experiment. The finite semantic space comprised 200 predicate-argument structures (excluding reflexivity), composed of 5 predicates (transitive verbs) and 5 individuals expressing external events, the binary CVs (200 = 5*5*4*2). The learner was exposed to half of the input data (100/200) as bottleneck. We performed experiments with varying degrees of compositionality of the initial language and observed the transition of topological similarity (TopSim), as a measure of compositionality (Brighton, Smith, & Kirby, 2005). We found that the high TopSim did not necessarily ensure the stable transmission of compositional conceptualized language. We also observed a sudden accidental decline of TopSim at certain generation, where agent had linguistic knowledge with weak expressivity and multiple distinct rules for a single meaning-form pair. This means that even if similar meanings, that is, one different element in predicate-argument structures, are mapped to similar forms, the production processes may largely differ. TopSim cannot represent this difference, indicating an issue in it.

3. Discussion

Humans conceptualize objective events from their own perspectives and reflect their conceptualized meanings in forms. Compositionality may facilitate learner or listener to infer the mapping between speaker's subjective conceptualized meaning and linguistic form. Thus, compositionality may work as a scaffold for evolving language that allows for communicating conceptualized meaning. However, as our result shows, a high degree of compositionality is not sufficient for transmission of a linguistic system with conceptualized meanings. Additionally, we point out the problem of TopSim in terms of conceptualized meaning. TopSim represents the correlation between the similarity among the objective meanings and that among the linear forms. Using TopSim as a compositionality measure presents at least two problems. One is that the calculation of TopSim does not consider the conceptualized meaning, so we here had to calculate it for each CVs separately. The second is that it assumes that in compositional languages similar forms correspond to similar meanings. However, when conceptualized meanings are reflected to forms in a language, as the change of linear order are often observed in natural languages, two different forms correspond to a single objective meaning, which causes decline in TopSim even if the language is compositional.

Acknowledgements

We gratefully thank the three anonymous reviewers for their invaluable critiques and insightful comments, leading to major revision. This work is partly supported by the Evolinguistics Foundation.

References

Brighton, H., Smith, K., & Kirby, S. (2005). Language as an evolutionary system. *Physics of Life Reviews*, *2*(3), 177–226.

Kirby, S. (2002). Learning, bottlenecks and the evolution of recursive syntax. In T. Briscoe (Ed.), *Linguistic evolution through language acquisition* (p. 173–204). Cambridge University Press.

Langacker, R. W. (2001). Discourse in cognitive grammar. *Cognitive Linguistics, 12*(2), 143–188.

Langacker, R. W. (2002). *Concept, image, and symbol*. Walter de Gruyter Inc.

Langacker, R. W. (2014). Conceptualization, symbolization, and grammar. In *The new psychology of language* (pp. 1-37). Psychology Press.

A tool for exploring building blocks in speech and animal vocalizations

Yannick Jadoul[*1,2] and Bart de Boer[2]

[*]Corresponding Author: Yannick.Jadoul@uniroma1.it
[1]Department of Human Neurosciences, Sapienza University of Rome, Rome, Italy
[2]Artificial Intelligence Lab, Vrije Universiteit Brussel, Brussels, Belgium

An important characteristic of human language is that it is based on combinatorial structure: it combines a small set of (mostly meaningless) building blocks into an unlimited set of (meaningful) utterances (de Boer, Sandler, & Kirby, 2012), which then in turn can get recombined into even bigger ones. Similar structure can be found in animal vocalizations (e.g., Kershenbaum et al., 2014) but the evolutionary link between say, primate vocalizations and human speech is not clear yet. Here we argue that to address this question we need tools to identify building blocks directly from the signal, but that these tools should be as little biased by linguistic notions as possible. Current analyses sometimes start from notions such as vowels (e.g., Fitch, de Boer, Mathur, & Ghazanfar, 2016; Boë et al., 2017) or consonants (e.g., Lameira, Maddieson, & Zuberbühler, 2014). Many other approaches assume a full segmentation of the input signal, and with no overlap or gaps between segments (e.g., Kreuk, Keshet, & Adi, 2020). We propose a tool, inspired by the field of data mining (frequent pattern mining; Aggarwal, Bhuiyan, & Hasan, 2014) to identify building blocks of speech directly from the signal, without assuming notions such as consonants, vowels, or syllables.

Identifying building blocks directly from a signal is hard, due to real signals being continuous and noisy. Even under ideal circumstances (single subject and low noise) no two occurrences of the same utterance are identical. We therefore adopt a number of techniques from speech recognition and combine them with frequent sequence mining to automatically derive candidate building blocks.

In order to demonstrate and evaluate our methods, we recorded a dataset with three-syllable nonsense words. Each CV-syllable consists of one of three consonants (/b/, /d/, /g/) and one of three vowels (/a/, /i/, /u/), resulting in 9 different syllables combined into $9^3 = 729$ highly structured nonsense words. All 729 words were read by a single speaker and recorded in a low-noise environment.

Frequent sequence mining algorithms can identify frequently occurring patterns in large sets of sequences. As they typically operate on symbolic sequences, a first step is to extract feature vectors from the acoustic data and cluster these in a set of discrete categories. Following standard speech processing procedure, we extracted 13 mel-frequency ceptral coefficients (MFCCs) using Parselmouth (Jadoul, Thompson, & de Boer, 2018; Boersma & Weenink, 2021) and used k-means to cluster the normalized MFCC vectors into 24 clusters. It should be noted that MFCCs are based on properties of human hearing, so they may need to be replaced by an appropriate model when analysing other species – based on the properties of that species' perception. This can easily be accommodated in our system.

The large number of patterns found by a frequent pattern mining algorithm need to be filtered before they can be interpreted as building blocks. On our small data set, running the CM-SPAM algorithm (Fournier-Viger, Gomariz, Campos, & Thomas, 2014; Fournier-Viger et al., 2016) already results in 7177 patterns that occur in more than 10% of input sequences. To filter these patterns to a manageable number of building blocks, we incrementally selected patterns which together cover more and more of the input sequences (Figure 1). The patterns we recover represent vowel formant patterns and formant transitions related to consonants.

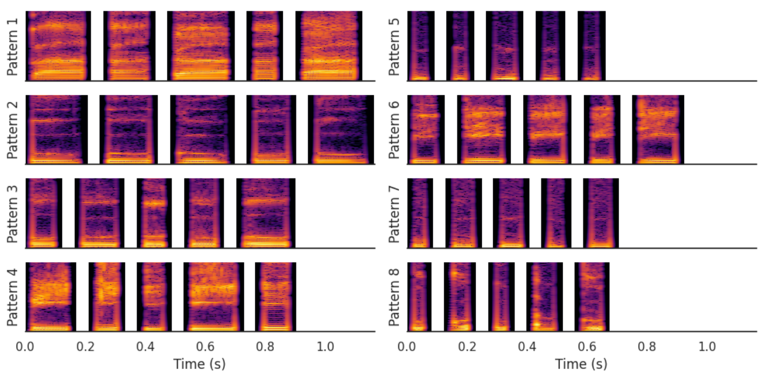


Figure 1. Audio fragments matched by the 8 most important building blocks show vowels and formant transitions representing consonants. The spectrograms' frequency ranges from 0 to 5000 Hz.

The current dataset is highly structured and has less noise than most real-world data, so it remains to be seen how our technique performs in less ideal circumstances. However, this preliminary result shows that our very general technique can identify relevant (i.e. in this case clearly related to the vowels and consonants in the original data) building blocks from a real signal, without assuming linguistic notions. Therefore it seems a promising approach for analysing the combinatorial structure in animal vocalizations, which could assist in further investigations into how the ability to use such structure has evolved.

References

Aggarwal, C. C., Bhuiyan, M. A., & Hasan, M. A. (2014). *Frequent pattern mining algorithms: A survey.* Springer.

Boersma, P., & Weenink, D. (2021). *Praat: doing phonetics by computer [Computer program].* Version 6.1.38, retrieved 2 January 2021 `http://www.praat.org/`.

Boë, L.-J., Berthommier, F., Legou, T., Captier, G., Kemp, C., Sawallis, T. R., Becker, Y., Rey, A., & Fagot, J. (2017). Evidence of a vocalic proto-system in the baboon (papio papio) suggests pre-hominin speech precursors. *12*(1), e0169321.

de Boer, B., Sandler, W., & Kirby, S. (2012). New perspectives on duality of patterning: Introduction to the special issue. *4*(4), 251–259.

Fitch, W. T., de Boer, B., Mathur, N., & Ghazanfar, A. A. (2016). Monkey vocal tracts are speech-ready. *2*(12).

Fournier-Viger, P., Gomariz, A., Campos, M., & Thomas, R. (2014). Fast vertical mining of sequential patterns using co-occurrence information. In *Advances in knowledge discovery and data mining: 18th pacific-asia conference, pakdd 2014, tainan, taiwan, may 13-16, 2014. proceedings, part i 18* (pp. 40–52).

Fournier-Viger, P., Lin, J. C.-W., Gomariz, A., Gueniche, T., Soltani, A., Deng, Z., & Lam, H. T. (2016). The spmf open-source data mining library version 2. In *Machine learning and knowledge discovery in databases: European conference, ecml pkdd 2016, riva del garda, italy, september 19-23, 2016, proceedings, part iii 16* (pp. 36–40).

Jadoul, Y., Thompson, B., & de Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, *71*, 1–15.

Kershenbaum, A., Bowles, A. E., Freeberg, T. M., Jin, D. Z., Lameira, A. R., & Bohn, K. (2014). Animal vocal sequences: not the markov chains we thought they were. *281*(1792), 20141370.

Kreuk, F., Keshet, J., & Adi, Y. (2020). Self-supervised contrastive learning for unsupervised phoneme segmentation. *arXiv preprint arXiv:2007.13465*.

Lameira, A. R., Maddieson, I., & Zuberbühler, K. (2014). Primate feedstock for the evolution of consonants. *18*(2), 60–62.

Multilevel phylogenetic inference of harmony in Indo-European

Yingqi Jing[1], Joakim Nivre[2], and Michael Dunn[3]

*Corresponding Author: yingqi.jing@lingfil.uu.se
*Department of Linguistics and Philology, Uppsala University, Uppsala, Sweden

One of the most well-known typological generalizations is that languages tend to order the grammatical head and its dependents in a consistent way (Greenberg, 1963; Dryer, 1992). For example, VO languages are more likely to have prepositions while OV languages are more likely to have postpositions. Over the past decades, new empirical findings and competing theories have been advanced to explain the word order harmony. These theories range from functional explanations focusing on cognitive and learning biases (Hawkins, 1983; Culbertson, Smolensky, & Legendre, 2012; Futrell, Levy, & Gibson, 2020), to alternative views, emphasizing the roles of cultural evolution and historical accidents in language change (Bybee, 1988; Dunn, Greenhill, Levinson, & Gray, 2011; Cristofaro, 2019). So far, there is little evolutionary evidence favoring or against harmony based on cross-linguistic corpus data, and it remains unclear whether the functional explanations can be reflected in the history of languages.

Using 45 dependency-annotated corpora from Universal Dependencies v2.10 (Zeman et al., 2022), we are trying to detect the evolutionary bias of harmony vs. disharmony in the history of Indo-European. To assess the cognitive benefits of consistency in head direction, we measure harmony by counting pairs of word orders (V-O and N-Gen, V-S and N-Adj, etc) that co-occur in the same direction in a sentence. Since word orders are not uniformly distributed, e.g., subjects almost always come before the verb, we need to control for the base distribution of each word order in a language. For this, we introduce two random baselines: one fixes the overall head direction in a language (random 1), and the other keeps unchanged the order of each dependency type in a language (random 2). By comparing the observed against the baselines, we can remove the confound of other possible processes (e.g., word order rigidity) in language change. We go beyond previous phylogenetic approaches that model the correlated evolution separately for each pair of word orders (Dunn et al., 2011; Jäger & Wahle, 2021). Instead, we have developed a multilevel phylogenetic Continuous-time Markov Chain model that can estimate evolutionary rates for harmony and disharmony at both population and group levels (Nalborczyk, Batailler, Lœvenbruck, Vilain, & Bürkner, 2019; Stan Development Team, 2022).

Our results reveal no clear difference in the estimated rate ratio for harmony between observed and random baselines (Figure 1). In particular, the observed rate ratio (mean rate ratio: 1.64, 90% CI = $[0.17, 4.47]$) has substantial overlaps with the second baseline (mean rate ratio: 1.76, 90% CI = $[0.17, 5]$). Our findings challenge the functional motivations for harmony during language comprehension, production, or learning. When the distribution of each individual word order is kept constant in a language, there is not much room left for any additional harmonic constraint between pairs of word orders in real utterances. This further suggests that the attested word order universals in previous work might emerge as a side-effect of word order rigidity in language evolution, and no appeal to cross-category harmony may be needed. In addition, when compared to the first baseline (mean rate ratio: 3.13, 90% CI = $[0.67, 6.23]$), the observed data show a lower rate ratio or a weaker evolutionary bias for harmony. This also contradicts previous theories that predict a general head-initial or head-final preference (Hawkins, 1994; Cancho, 2017). Conversely, word orders seem to be less constrained than commonly assumed, and they tend to evolve towards a more mixed configuration at least in Indo-European. Further research is needed by extending the approach to other families before we can draw firm conclusions.

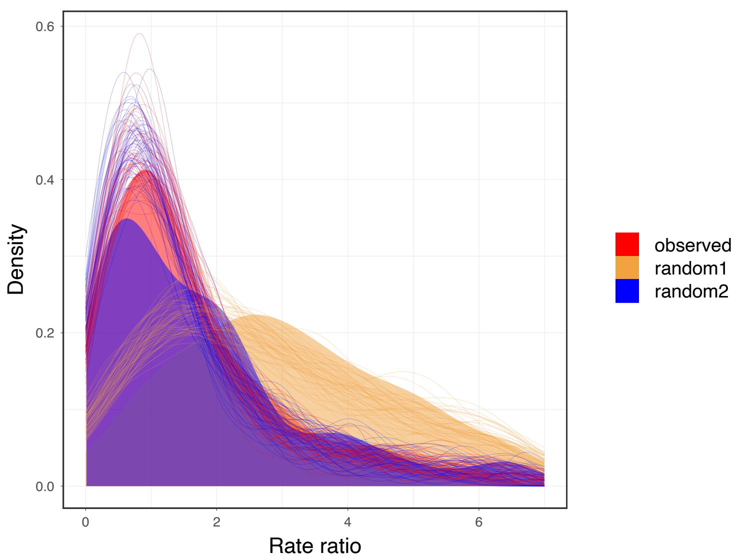


Figure 1. Posterior density of rate ratio of harmony to disharmony from the multilevel phylogenetic model. Higher rate ratios indicate a stronger evolutionary bias towards harmony. The shaded areas under the curve represent the rate ratio at the population level ("fixed effects"), and the thin lines represent the rate ratio at the group level ("random effects").

References

Bybee, J. L. (1988). The diachronic dimension in explanation. In J. A. Hawkins (Ed.), *Explaining language universals* (p. 350-379). Basil Blackwell.

Cancho, R. Ferrer i. (2017). The placement of the head that maximizes predictability. an information theoretic approach. *Glottometrics*, *39*, 38–71.

Cristofaro, S. (2019). Some language universals are historical accidents. In S. Ilja A., M. Susanne, L. Natalia, & S.-B. Karsten (Eds.), *Taking diachronic evidence seriously: Result-oriented vs. source-oriented explanations of typological universals* (Vol. 63, p. 25-46). Language Science Press.

Culbertson, J., Smolensky, P., & Legendre, G. (2012). Learning biases predict a word order universal. *Cognition*, *122*(3), 306–329.

Dryer, M. S. (1992). The greenbergian word order correlations. *Language*, *68*(1), 81–138.

Dunn, M., Greenhill, S. J., Levinson, S. C., & Gray, R. D. (2011). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, *473*(7345), 79–82.

Futrell, R., Levy, R., & Gibson, E. (2020). Dependency locality as an explanatory principle for word order. *Language*, *96*(2), 371-412.

Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In J. H. Greenberg (Ed.), *Universals of language* (p. 58-60). MIT Press.

Hawkins, J. A. (1983). *Word order universals.* New York: Academic Press.

Hawkins, J. A. (1994). *A performance theory of order and constituency.* Cambridge: Cambridge University Press.

Jäger, G., & Wahle, J. (2021). Phylogenetic Typology. *Frontiers in Psychology*, *12*.

Nalborczyk, L., Batailler, C., Lœvenbruck, H., Vilain, A., & Bürkner, P.-C. (2019). An introduction to bayesian multilevel models using brms: A case study of gender effects on vowel variability in standard indonesian. *Journal of Speech, Language, and Hearing Research*, *62*(5), 1225–1242.

Stan Development Team. (2022). *RStan: the R interface to Stan.* (R package version 2.26.13)

Zeman, D., et al.. (2022). *Universal dependencies 2.10.* (LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University)

Simulating the spread and development of protolanguages

Sverker Johansson*[1]

*Corresponding Author: sja@du.se
[1]Dalarna University, Falun, Sweden

Simulation of language change can (1) provide training and validation data for language reconstruction, (2) explore protolanguage hypotheses where real data are unavailable, and (3) test the importance of different processes in language evolution. But previous simulation work is mainly either micro-scale or too abstract. Neither pole captures the full range of language dynamics. The aim here is to fill that gap between micro and macro. The simulation combines explicit models for language, demography, and geography in sufficient detail to produce output that in relevant respects mimics real language data, with adequate scale and scope to model language history across millennia.

1. Why simulation?

Languages change over time, but our understanding of the relative importance of different processes in the distant past remains limited. The development of methods for reconstructing language change is also hampered by a shortage of suitable data.

Simulating language change in software can help alleviate these problems (cf. List, 2019). Virtually unlimited amounts of simulated language data can be produced where the processes are known and controllable, and the true diversification path is known.

Furthermore, tuning process strength in simulation until the results resemble real language diversity may inform theories of language dynamics, within the limits set by the problem of equifinality (Kandler & Powell, 2018). Early forms of protolanguages may be simulated by adding suitable constraints to the simulation (cf. Gong et al., 2022).

But simulated data will only be helpful if the simulation reproduces relevant aspects of reality closely enough. Several items in List (2019) *Open problems in* *computational linguistics* concern simulation issues. Extant simulations are mainly of two types:

- Detailed short-term simulations of dynamics within a single language, often agent-based (e.g. Cangelosi & Parisi, 2002; Nolfi & Mirolli, 2010, as well as many Evolang entries over the years, e.g. Wang & Steels, 2008).
- Macro-scale long-term simulations that cover the dynamics between languages, but with linguistic and/or geographical details abstracted away (e.g. Hochmuth et al., 2008; Wichmann, 2017, 2021; Kapur & Rogers, 2020; Ciobanu & Dinu, 2018; Gergel et al., 2021).

Neither type covers the middle ground where within-language and between-languages dynamics meet. This work aims to fill that gap, with a simulation that has sufficient linguistic, geographic and anthropological detail to produce data that are useful for a range of purposes, and sufficient scope to cover macro-scale dynamics over millennia.

This simulation is not tailored to test specific hypotheses, but to produce simulated language data that is useful for further processing and testing.

2. Simulation framework

The simulation contains the following core models:

- Explicit *geography* model.
  - Topography, climate, vegetation.
  - Climate change.
- Explicit *population* model.
  - Population growth and decline.
  - Migration & population split.
  - Population interactions.
- Explicit *language* model.
  - Within-language processes.
  - Between-language dynamics.
- Technological development.

The basic simulation unit here is a speech community with typically 100-1000 speakers, speaking a common language. To initialize the simulation, real languages are used as seed languages, which then evolve through regular sound change, word gain and loss, semantic shift, language contact, and areal effects. All processes are adjustable and can be disabled. Lexical data for seed languages are taken from CLICS3 (Rzymski et al., 2019), and grammatical data from Grambank (Skirgård et al., 2023).

The geography of the real world is used, with topography from De Ferranti (2015), rivers from Kelso (2016), climate/ecology from NASA (2016) and climate change from Snyder (2016). Each speech community lives in a grid square (default size 50x50 km, which is also the resolution of the geographic model) which may be shared with other communities up to a carrying capacity. The carrying capacity depends on climate, vegetation, and access to water, and may fluctuate from year to year (modelling drought etc.).

The population of each community may increase or decrease over time, depending on food availability, and surplus population may migrate to greener pastures, forming a new community. The new community speaks a clone of the parent community language, but their languages then evolve independently. The distance travelled in migration depends on real terrain and available technology.

Technological innovations occur occasionally, starting from a paleolithic level. An innovation may increase food production in some or all environments, open up new environments to exploitation, or enhance mobility. Some innovations are prerequisites for others. One community may learn a technology from a neighboring community. The technological model reaches Bronze Age technology, and is inspired by the computer game Civilization (Meier, 2021). This allows communities to evolve from hunter-gatherers to horticulturalists, pastoralists, and farmers.

3. Language model

Each language in the simulation has an explicit vocabulary. The word forms are strings of phonemes, each paired with one or more meanings. Each meaning may be covered by zero, one, or more words. Grammar is modelled using typological parameters from GramBank (Skirgård et al., 2023).

New languages are born when a community splits in two. Languages can die in two different ways: the whole community may starve, or the community may be assimilated into a more powerful neighboring community.

The processes affecting a language can be divided into *endogenous* processes internal to it, and *exogenous* processes due to contact with other languages.

3.1 Endogenous processes

Regular sound change is modelled as one random phoneme in the language being replaced by another phoneme with similar features, at random points in time at some rate. Sound change may be either unconditional or conditional. Words may undergo metathesis, swapping two nearby phonemes in the word. Other sound change processes are not implemented yet.

Semantic shifts are implemented using colexification data from CLICS3 (Rzymski et al., 2019). The meaning of a word may be broadened to include another meaning, with a rate proportional to the colexification rate between the original and the new meaning. A word may also lose semantic scope, especially if it has synonyms in one of its senses.

Grammatical change is modelled as random changes in typological parameters, with care taken to keep the resulting grammar consistent.

3.2 Exogenous processes

Languages that are in contact regularly borrow words from each other. Terrain and travel technology affect which languages are regarded as in contact. Long-range borrowing beyond the regular travel range happens occasionally. Borrowing rate is enhanced in the following cases:

- Two languages occupy the same grid square.
- Two languages are closely related (either short time since split or short lexical distance between vocabularies).
- One language lacks a word for a concept.

When technology is transferred between communities, the vocabulary for the new technology is also borrowed.

If multiple languages are present in the same grid square, minorities are heavily affected by the most powerful group. For each generation, some fraction of the minority will shift to the majority language.

3.3 Areal effects

Areal effects are exogenous processes, where language features spread broadly in a region so that unrelated neighboring languages come to resemble each other. This is modelled for sounds, words, and grammatical features in roughly the same way: if a large fraction of the languages in a region share the same feature, the languages that don't have it are likely to adopt it.

4. Flexibility

All processes in the model can be switched on and off under user control, either through runtime switches or through parameter files. All processes can also have their rates adjusted through parameter files. The parameter space is by necessity unexplorably large in any detailed simulation (30-odd primary parameters in this case, plus huge transition matrices), but having the parameters visible instead of fixed in code makes this issue explicit. Both time step and total running time for the simulation can be chosen at runtime.

A smaller geographical region than the whole world can be chosen at runtime, in order to run contact-rich scenarios. It is also possible to create and load alternative geography data, in order to test hypotheses about the importance of geographical structure.

Seed languages can either be selected from a list, or a random selection of a given number of languages can be provided.

5. Output

Simulation results are available in several different formats:

- Language metadata: seed, birth year, birth place, death year
- Swadesh matrices (tab-separated text; can be read by e.g. MS Excel).
- Word lists in CLDF format.
- Grammar data as list of typological features.
- True phylogenetic tree in NEXUS format
- True cognate lists

Simulation results and the underlying true data are consistently saved in separate files, that can be cross-referenced with unique identifiers.

6. Validation

The aim of the simulation is to produce data that mimic patterns in real language data in relevant respects. On visual inspection, the output generally looks plausible, though with some unusual phoneme sequences; phonotactic constraints are not implemented. But visual impression is of course not a scientific validation.

Comparing statistical patterns in the output with real language data is a more reliable validation method than visual impressions. In Figure 1 below is one example. The similarity between word forms is quantified using weighted Levenshtein (1966) distances. Between unrelated words, the distances should be randomly distributed, whereas cognates can be expected to have smaller distances. In each part of Figure 1, the dashed curve is for words from different language families, and the solid curve is for words with the same meaning within the same language family. The left diagram shows real data from CLICS3, the right is simulated data. The two diagrams are qualitatively similar but not identical; part of the difference may be due to the simulated families all having the same age. If the simulation is left running long enough (15,000+ years) the two curves will eventually converge, as cognates are no longer discernible.

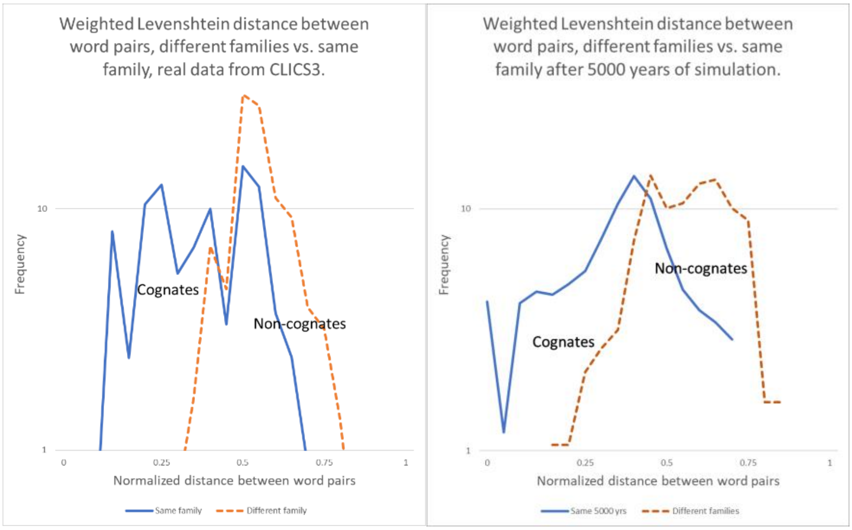


Figure 1. Levenshtein distance between word pairs, comparison between real and simulated data.

Another validation option is to use the output from the simulation as input to the same phylogenetic reconstruction methods that are used in computational historical linguistics with real data (cf. Jäger, 2019). This works fine, the standard program PAUP (Swofford, 1996) for phylogenetic reconstruction recovers the true tree with a plausible level of accuracy. Automated cognate detection likewise recovers the true cognate sets from simulated data with reasonable accuracy.

7. Summary

The simulation basically works as intended, producing reasonable-looking data in large quantities. Some tuning work is still needed, but the model passes basic validation tests. The output works as input to phylogenetic reconstruction.

The software runs on a regular PC, generating thousands of languages over thousands of years in a matter of hours. But over very long time scales with very large numbers of languages, it will bog down computationally.

8. Supplementary Materials

Software and sample output available at
https://github.com/Lsjbot/LangChangeSimulator/tree/master

References

Cangelosi, A. & Parisi, D. (2002) *Simulating the evolution of language.* London: Springer.

Ciobanu, A. M. & Dinu, L. P. (2018) Simulating language evolution: A tool for historical linguistics. *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 68–72 Santa Fe, New Mexico, USA, August 20-26, 2018.

De Ferranti, J. (2015) *Viewfinder Panoramas Digital Elevation Model.* http://www.viewfinderpanoramas.org/dem3.html

Dellert, Johannes. (2019) *Information-theoretic causal inference of lexical flow (Language Variation 4).* Berlin: Language Science Press.

Gergel, R., Kopf-Giammanco, M., & Puhl, M. (2021). Simulating semantic change: A methodological note. *Experiments in Linguistic Meaning, 1*, 184-196.

Gong, T., Shuai, L. & Yang, X. (2022). A simulation on coevolution between language and multiple cognitive abilities. *J Lang Evo* 2022:120-145

Hochmuth, M., Lüdeling, A., & Leser, U. (2008). *Simulating and reconstructing language change*. Unpublished manuscript, Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät II, Institut für Informatik. https://www.informatik.hu-berlin.de/de/forschung/gebiete/wbi/research/publications/2008/tr_language_change.pdf

Jäger, G. (2019) Computational historical linguistics. *Theoretical Linguistics*, 45(3-4):151-182. https://doi.org/10.1515/tl-2019-0011

Kandler, A. & Powell, A. (2018) Generative inference for cultural evolution. *Phil. Trans. R. Soc. B*373:201700562.0170056 http://doi.org/10.1098/rstb.2017.0056

Kapur, R & Rogers, P (2020) *Modeling language evolution and feature dynamics in a realistic geographic environment.* In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona.

Kelso, N V (2016) *Natural Earth Data*. https://www.naturalearthdata.com/downloads/

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

List, Johann-Mattis (2019): *Open problems in computational historical linguistics.* Invited talk presented at the 24th International Conference of Historical Linguistics (2019-07-01/05, Canberra, Australian National University).

Meier, Sid (2021) *Sid Meier's Memoir.* W W Norton.

NASA (2016) *NASA Earth Observations.* https://neo.gsfc.nasa.gov/

Nolfi, S & Mirolli, M (2010) *Evolution of Communication and Language in Embodied Agents.* Springer.

Rzymski, Christoph and Tresoldi, Tiago et al. (2019). *The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies*. DOI: 10.1038/s41597-019-0341-x

Skirgård et al. (2023). Grambank v1.0 (v1.0.3) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7844558

Snyder, C. (2016) Evolution of global temperature over the past two million years. *Nature* **538**, 226–228. https://doi.org/10.1038/nature19798

Swofford, D. (1996). *PAUP*: Phylogenetic analysis using parsimony (and other methods), version 4.0*. Sunderland, MA: Sinauer Assoc.

Wang, E. & Steels, L. (2008) Self-interested agents can bootstrap symbolic communication if they punish cheaters, In Proceedings of Evolang 7.

Wichmann, S. (2017) Modeling language family expansions. *Diachronica 34:1, 79-101*.

Wichmann, S. (2021) A world language family simulation. *Physica A: Statistical Mechanics and its Applications*, 572: 125913. https://doi.org/10.1016/j.physa.2021.125913 .

Emergent grammar from a minimal cognitive architecture

Anna Jon-And[*1,2] and Jérôme Michaud[1,3]

[*]Corresponding Author: anna.jon-and@su.se
[1]Centre for Cultural Evolution, Department of Psychology, Stockholm University, Sweden
[2]Department of Romance Studies and Classics, Stockholm University, Sweden
[3]Department of Mathematics and Physics, Mälardalen University, Västerås, Sweden

In this paper, we introduce a minimal cognitive architecture designed to explore the mechanisms underlying human language learning abilities. Our model inspired by research in artificial intelligence incorporates sequence memory, chunking and schematizing as key domain-general cognitive mechanisms. It combines an emergentist approach with the generativist theory of type systems. By modifying the type system to operationalize theories on usage-based learning and emergent grammar, we build a bridge between theoretical paradigms that are usually considered incompatible. Using a minimal error-correction reinforcement learning approach, we show that our model is able to extract functional grammatical systems from limited exposure to small artificial languages. Our results challenge the need for complex predispositions for language and offer a promising path for further development in understanding cognitive prerequisites for language and the emergence of grammar during learning.

1. Introduction

The question of what cognitive mechanisms underlie human language learning abilities is a long-standing controversy. Theories on inborn language organisation (Chomsky, 1957; Pinker & Jackendoff, 2005; Fodor, 1983) or learning biases (Nowak, Komarova, & Niyogi, 2002; Reali & Griffiths, 2009; Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010; Tenenbaum, Kemp, Griffiths, & Goodman, 2011) often disregard social aspects of human learning and the resulting variations in language structures. Furthermore, the claim that language learning requires predefined linguistic predispositions has been challenged by the achievements of modern language models (Piantadosi, 2023). In contrast, theories relying on domain-general learning with no specific predispositions for language (Bybee, 1985; Tomasello, 2003; Heyes, 2018) and culturally emergent structure (Kirby, Cornish, & Smith, 2008; Goldberg, 2007; Langacker, 1987; Croft, 2001) generally lack explicit suggestions on the machinery underlying such learning. Connectionist models have accounted for some aspects of language acquisition (Elman, 1996; Christiansen & Chater, 2001; McClelland et al., 2010; Piantasodi & Hill, 2022), but they are hard to interpret and face challenges in capturing symbolic representations.

Here, we suggest a simple formal operationalization of emergentist theories, grounded in cognitive architecture artificial intelligence research, where few general mechanisms should explain diverse and complex cognitive phenomena (Newell, 1994). To select relevant mechanisms, we start by considering minimal cognitive differences between humans and other animals. The empirically well supported sequence hypothesis postulates that faithful perception of order is uniquely human and a defining feature of human cognition (Grant, 1976; MacDonald, 1993; Ghirlanda, Lind, & Enquist, 2017; Read, Manrique, & Walker, 2021; Lind, Vinken, Jonsson, Ghirlanda, & Enquist, 2023; Enquist, Ghirlanda, & Lind, 2023; Jon-And, Jonsson, Lind, Ghirlanda, & Enquist, 2022). Processing of language or any sequential information also requires a capacity for chunking, i.e. considering a recurrent sequence of stimuli as a unit (Tomasello, 2003; Bybee, 2002; Servan-Schreiber & Anderson, 1990; McCauley & Christiansen, 2019; Christiansen & Chater, 2016; Miller, 1956; Cowan, 2001). The third feature of our model is schematizing, a fundamental aspect of learning in humans and other animals (Hull, 1943; Ghirlanda & Enquist, 2003), that is required for categories to emerge. With these three components, we build a minimal reinforcement learning model aiming at accounting for the emergence of grammatical categories from limited exposure to small artificial languages.

2. The learning task

During language learning words are perceived in a stream and the learner's task is to identify sentences. While research has focused on word segmentation (Saffran, Aslin, & Newport, 1996; Saffran, 2001), higher levels of segmentation are also necessary for language understanding. Here, we test the hypothesis that grammar emerges to facilitate language processing at the sentence level. Even though real life language learners receive support for segmentation from, for instance, prosodic cues (Kuhl, 2004), this support is absent in our model for reasons of simplicity and feasibility, and enables investigating how far the system can reach without it. We assume that the identification of meaningful units triggers internal or external rewards, as it contributes to understanding, even though no explicit instructions or feedback are present. Following Sutton & Barto (2018), we define the task as a Markov Decision Process (MDP): a framework for studying learning from interactions with an environment to achieve a goal. In our case, the goal is to identify sentences in a stream of words constituting the environment. Sentences are generated using probabilistic context free grammars. Words' frequency is inversely proportional to their rank, approximating distributions in natural languages (Zipf, 1932). Any hint as to the beginning of sentences, for example capital letters or punctuation are removed from the list of words. For example,

The cat chases the dog. The man loves his girlfriend. The sun shines.

will be transformed into

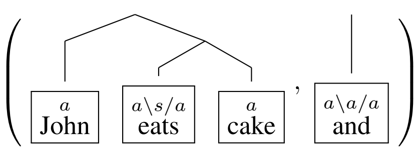
the cat chases the dog the man loves his girlfriend the sun shines

and the goal is to correctly identify sentences boundaries. The information about sentence boundaries is thus masked and used to drive the agent's learning. To assist the segmentation task, we implement schematizing using an adaptation of Lambek's syntactic type theory (Lambek, 1958), which provides the simplest mathematical framework for handling order, composition and abstraction of words (Heunen, Sadrzadeh, & Grefenstette, 2013). In this theory, every word is assigned a symbolic formula encoding how it can be grouped with other words. Primitive types are represented by one symbol, e.g. a or b, and compound types are represented as a/b or $b\backslash a$ and can be thought of as production rules encoding how types can be grouped to generate higher order types. For example, a/b followed by b or b followed by $b\backslash a$ both reduce to a. In the example

John	works	here
n	$n\backslash s$	$s\backslash s$

John is assigned the type n and *works* is assigned the compound type $n\backslash s$. Grouping them leads to typing the chunk *John works* as s for sentence. Furthermore, *here*, typed as $s\backslash s$, transforms the sentence *John works* into the longer sentence *John works here*. Our adaptation of the theory assumes no predefined categories.

In an MDP, the agent starts in a state, takes actions to move to a new state, and gets rewards (positive or negative) to drive the learning process. In our model, the possible states are encoded as pairs of structured *chunks*. A chunk refers to a sequence of potentially typed words and an associated binary tree with the words as leaves. By default, the second and most recently perceived chunk always consists of a single word. An example state is given by:



From a given state, and following an incremental processing of input akin to a chunk and pass mechanism (Christiansen & Chater, 2016), the learner needs to decide whether to insert the second element in the currently analyzed structure or to place a boundary between the two chunks. The possible resulting states corresponding to our example are displayed in Figure 1. Cases (1), (2), and (3) correspond to chunking actions in which the second element is inserted at different levels in the tree and case (4) corresponds to boundary placement, which triggers reinforcement. Positive rewards are given whenever the identified sentence is correct and negative rewards are given otherwise. We note that in this example, if previous types are correctly assigned, only case (1) is compatible with

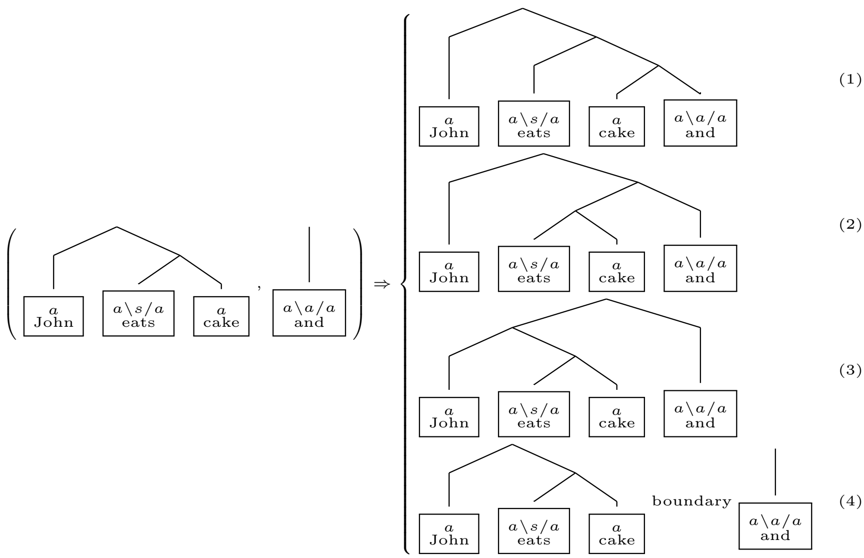


Figure 1. Possible decisions in a given state.

the assigned types since the list of types $[a, [a\backslash s/a, [a, a\backslash a/a]]]$ reduces to s/a. Placing a boundary is also incompatible with the assigned types because while the first element correctly reduces to a sentence, the second element expects to be preceded by a. The primary goal of the model is to successfully segment sentences and the secondary goal is to assign syntactic types to words to induce a grammatical representation of the input.

3. The cognitive architecture and learning mechanism

The cognitive architecture consists of a long term memory storing associations between states and actions as well as associations between words and types and a working memory used to process sentences, to invent types, and to update the long term memory through reinforcement learning.

Sentence processing works as follows. The learner initializes its state by reading the first two words in the stream and try to assign syntactic types to them. In the beginning of the learning process, no types are associated with words, but types will emerge during the learning process. If the words have compatible types, these types are chosen, otherwise, if only one word has a type, it assigns a compatible type to the other word. Once the typing process is done, actions are chosen based on both the state-action associations and the assigned types, if any. The process continues until the chosen action is a boundary placement. In that case, reinforcement is triggered. This is done using an error-correction algorithm (Rescorla

& Wagner, 1972; Sutton & Barto, 2018) that exploits the hierarchical structure of the states and takes into account whether type associations have informed the decision or not. Upon correct boundary placement, state-action associations and word-type associations are positively reinforced. If the words in the sentence were not typed, the recognized structure is typed s and compatible types are assigned to individual words constituting the sentence, randomly choosing among possible assignments. Initial types are always constructed minimalistically containing only the primitive s. Upon incorrect boundary placement, state-action associations and word-type associations are negatively reinforced. When all reinforcements have been performed, the working memory is emptied and a new sentence is processed.

During sentence processing, word-type associations that have decreased below a given threshold block type assignment and trigger the emergence of new primitive types, providing a mechanism to invent new types. Differently from the original theory of types and other categorial-based formalisms (Steedman & Baldridge, 2011; Kogkalidis, Moortgat, & Moot, 2020), we do not assume any predefined grammatical categories. Types are invented when needed, replicated through type assignment and selected based on their usefulness in sentence segmentation. These general processes of invention, replication and selection turn our type system into an evolutionary system following a broad definition of evolution (Hull, 2001). We call this modified framework an evolutionary type system.

4. Results and discussion

The evaluation of the cognitive architecture's ability to extract grammar from small artificial languages comprises two aspects: (i) do types emerge that make learning faster than learning without types (i.e. a model version without schematizing); (ii) do the types that emerge expose any resemblance with grammatical categories? If these two criteria are fulfilled, results support the hypothesis that grammar emerges because it makes learning more efficient.

We have implemented a pilot type system for two-word sentences, consisting of nouns and intransitive verbs. A more general type system that will encompass longer sentences like those previously exemplified is currently being developed.

Table 1. Stimulus-type associations (between -2 and 10) for a language with 4 nouns and 4 verbs.

	Noun 1	Noun 2	Noun 3	Noun 4	Verb 1	Verb 2	Verb 3	Verb 4
a	10	10	10	10				
s	-0.1	-0.5	-0.5	0.6				
$a\backslash s$					10	10	10	10
$s\backslash s$						1	1	
s/a					-0.38	-0.2	-0.4	

We see that functional types emerge. The left panel of Fig. 2 displays an example

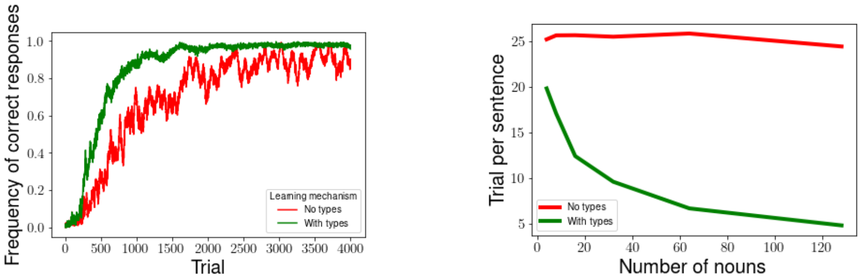


Figure 2. Left: Learning curves for a language with 16 nouns and 8 intransitive verbs with and without types. Right: Learning times for a varying number of nouns and a fixed number of verbs.

of a language with 16 nouns and 8 verbs, where learning to identify sentences is faster with types. The right panel shows that the difference in learning times increases with vocabulary size. For a learner without a type system, the number of trials it takes to identify each sentence is independent of vocabulary size, while for a learner with the evolutionary type system it decreases when the vocabulary grows, reflecting the productive assignment of types to novel words. This function is arguably similar to that of productive grammatical categories.

Table 1 shows final word-type association values. The strongest word-type associations correspond to nouns being typed as a and verbs being typed as $a\backslash s$. Note that the only type assigned to Verb 4 is $a\backslash s$, which indicates successful generalization. These emergent types are coherent with the original theory of syntactic types (Lambek, 1958) and indicate that the second element is analyzed as the predicate or head of the sentence and the first element as an argument or a dependent. The emergent types are parsimonious since they only use two primitives and all nouns and verbs have the same respective strongest type. Despite the fact that type invention and assignment are symmetric, the head second order is dominant. This is likely a consequence of the learning process and the fact that it is easier to generate blocking types for the first word of the sentence that for the second one, leading to a higher likelihood of nouns being assigned a new primitive type. This order bias is compatible with the fact that subject-verb is a more frequent word order in the world's languages than verb-subject .

Our pilot results support the hypothesis that accurate sequence perception combined with chunking and schematizing may suffice for learning grammar, which would imply that previously suggested more complex predispositions for language are unnecessary. The results also suggest that grammar can emerge as a self-organizing solution to the combinatorial problem of language learning. These results are promising for further development of the model, extending its capacity to process sentences of any length and structure, and applications to natural language corpora.

Acknowledgements

This work was supported by the Swedish Research Council (VR 2022-02737) and by the Marianne and Marcus Wallenberg Foundation (grant no. 2021.0039).

References

Bybee, J. (2002). Phonological evidence for exemplar storage of multiword sequences. *Studies in second language acquisition*, *24*(2), 215–221.

Bybee, J. L. (1985). *Morphology: A study of the relation between meaning and form.* John Benjamins Publishing.

Chomsky, N. (1957). *Syntactic structures.* The Hague/Paris: Mouton.

Christiansen, M. H., & Chater, N. (2001). *Connectionist psycholinguistics.* Greenwood Publishing Group.

Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and brain sciences*, *39*.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences*, *24*(1), 87–114.

Croft, W. (2001). *Radical construction grammar: Syntactic theory in typological perspective.* Oxford University Press on Demand.

Elman, J. L. (1996). *Rethinking innateness: A connectionist perspective on development* (Vol. 10). MIT press.

Enquist, M., Ghirlanda, S., & Lind, J. (2023). *The human evolutionary transition: From animal intelligence to culture.* Princeton University Press.

Fodor, J. A. (1983). *The modularity of mind.* MIT press.

Ghirlanda, S., & Enquist, M. (2003). A century of generalization. *Animal Behaviour*, *66*(1), 15–36.

Ghirlanda, S., Lind, J., & Enquist, M. (2017). Memory for stimulus sequences: a divide between humans and other animals? *Open Science*, *4*(6), 161011.

Goldberg, A. (2007). *Constructions at work : the nature of generalization in language.* Oxford University Press.

Grant, D. S. (1976). Effect of sample presentation time on long-delay matching in the pigeon. *Learning and Motivation*, *7*(4), 580–590.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in cognitive sciences*, *14*(8), 357–364.

Heunen, C., Sadrzadeh, M., & Grefenstette, E. (2013). *Quantum physics and linguistics: a compositional, diagrammatic discourse.* Oxford University Press, USA.

Heyes, C. (2018). *Cognitive gadgets: the cultural evolution of thinking.* Harvard University Press.

Hull, C. L. (1943). Principles of behavior: an introduction to behavior theory.

Hull, D. L. (2001). *Science and selection: Essays on biological evolution and the philosophy of science.* Cambridge University Press.

Jon-And, A., Jonsson, M., Lind, J., Ghirlanda, S., & Enquist, M. (2022). Sequence representation as an early step in the evolution of language. *PsyArXiv*.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *PNAS*, *105*(31), 10681–10686.

Kogkalidis, K., Moortgat, M., & Moot, R. (2020). Neural proof nets. *arXiv preprint arXiv:2009.12702*.

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nature reviews neuroscience*, *5*(11), 831–843.

Lambek, J. (1958). The mathematics of sentence structure. *The American Mathematical Monthly*, *65*(3), 154–170.

Langacker, R. W. (1987). *Foundations of cognitive grammar: Volume i: Theoretical prerequisites* (Vol. 1). Stanford university press.

Lind, J., Vinken, V., Jonsson, M., Ghirlanda, S., & Enquist, M. (2023). A test of memory for stimulus sequences in great apes. *Plos one*, *18*(9), e0290546.

MacDonald, S. E. (1993). Delayed matching-to-successive-samples in pigeons: Short-term memory for item and order information. *Animal Learning & Behavior*, *21*(1), 59–67.

McCauley, S. M., & Christiansen, M. H. (2019). Language learning as language use: A cross-linguistic model of child language development. *Psychological review*, *126*(1), 1.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in cognitive sciences*, *14*(8), 348–356.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, *63*(2), 81.

Newell, A. (1994). *Unified theories of cognition.* Harvard University Press.

Nowak, M. A., Komarova, N. L., & Niyogi, P. (2002). Computational and evolutionary aspects of language. *Nature*, *417*(6889), 611-617.

Piantadosi, S. (2023). Modern language models refute chomsky's approach to language. *Lingbuzz Preprint, lingbuzz, 7180*.

Piantasodi, S. T., & Hill, F. (2022). Meaning without reference in large language models. *arXiv preprint arXiv:2208.02957*.

Pinker, S., & Jackendoff, R. (2005). The faculty of language: what's special about it? *Cognition*, *95*(2), 201–236.

Read, D. W., Manrique, H. M., & Walker, M. J. (2021). On the working memory of humans and great apes: Strikingly similar or remarkably different? *Neuroscience & Biobehavioral Reviews*.

Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, *111*(3), 317–328.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement [incollection]. In *Classical conditioning: current research and theory* (p. 64-69). Appleton-Century-Crofts.

Saffran, J. R. (2001). Words in a sea of sounds: The output of infant statistical learning. *Cognition*, *81*(2), 149–169.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Servan-Schreiber, E., & Anderson, J. R. (1990). Learning artificial grammars with competitive chunking. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(4), 592.

Steedman, M., & Baldridge, J. (2011). Combinatory categorial grammar. *Non-Transformational Syntax: Formal and Explicit Models of Grammar*, 181–224.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction.* MIT press.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *science*, *331*(6022), 1279–1285.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition.* Harvard University Press.

Zipf, G. K. (1932). *Selected studies of the principle of relative frequency in language.* Harvard university press.

Adapting to individual differences: An experimental study of variation in language evolution

Mathilde Josserand[*1,2], François Pellegrino[2], Oxana Grosseck[3], Dan Dediu[4,5], and Limor Raviv[3,6]

[*]Corresponding Author: mathilde.josserand@gmail.com

[1] Laboratoire Eco-Anthropologie UMR 7206, CNRS/MNHN/Université Paris Cité, Paris, France

[2] Laboratoire Dynamique du Langage, Université Lumière Lyon 2 - CNRS UMR 5596, Lyon, France

[3] LEADS group, Max Planck Institute for Psycholinguistics, Nijmegen, NL

[4] Department of Catalan Philology and General Linguistics, University of Barcelona, Barcelona, Spain

[5] University of Barcelona Institute for Complex Systems, Barcelona, Spain

[6] cSCAN, University of Glasgow, Glasgow, UK

While language variation has been typically studied as a marker of social groups (Labov, 1972, 1990), it is also omnipresent at the individual level (Yu & Zellou, 2019). Randomly pick two individuals in any population, and it is highly likely that they will exhibit some degree of variability in their language, for example by pronouncing the same word slightly differently. Language variation may arise from various causes, such as individual differences in the physiological ability to produce sounds. But could these individual differences shape the trajectory of language evolution?

Studies on language evolution traditionally disregard individual differences, and instead focus on group-level patterns and universal biases (despite a long tradition of studying variation in sociolinguistics, e.g. Tamminga, 2021, with which our work is connected, but different in the types of individual differences, processes and phenomena of interest). However, treating groups as homogeneous entities can result in a loss of valuable information when it comes to understanding language evolution. Indeed, languages could adapt to the biases of only a subset of their speakers (Butcher, 2018), as well as to the anatomical and physiological traits of their speakers (Blasi et al., 2019; Dediu et al., 2017, 2019). Individuals may unconsciously adapt their language to align with the unique characteristics of their conversational partners (i.e., accommodation), which can contribute to patterns of language evolution (Fehér et al., 2016, 2019). While some agent-based models have attempted to investigate language evolution dynamics in heterogeneous populations (Jameson & Komarova, 2009a, 2009b; Josserand et al., 2021; Navarro et al., 2018; Rita et al., 2022), no experimental work to date has tested the role of individual differences in shaping the live formation of languages in the lab.

In this study, we use an experimental approach based on a group communication game (Raviv et al., 2019a, 2019b), where micro-societies of four interacting participants need to communicate with each other using a miniature artificial language. Following exposure to a set of initial labels, participants interacted face-to-face in alternating dyads over multiple rounds using a computer interface. In each round, one participant produced a label to describe an image to their partner, and their partner needed to select the correct item from a set of distractors. We introduced individual differences by preventing one participant (the *biased participant*) from using two (out of eight) letters on the keyboard, simulating a speech variation or impairment that individuals may experience in their real lives. We ask whether the collective language of the group will adapt to the biased participant (i.e., by avoiding the use of these 'unavailable' letters), or, conversely, whether the unbiased majority will prevail (i.e., resulting in a language that encompasses all letters).

We tested 7 groups containing one biased participant (*heterogeneous* groups) and 7 groups containing no biased participants (*control* groups). Using mixed-effect models with the group as a random factor, we assessed participants' communicative success, their convergence on a shared language, and the specific usage frequency of the letters unavailable to the biased participants in the other participants' production in different models. Our results show that languages evolved differently in groups with a biased participant. After the nine rounds of communication, *heterogeneous* groups showed less communicative success (β=-27±7.3, $p<0.01$) and convergence (β=-0.33±0.04, $p<0.001$), which fostered the use of more linguistic variants. Additionally, we noticed partner-specific alignment, with participants typically adjusting their language to accommodate the idiosyncrasies of the biased participant during one-on-one interactions (i.e., avoiding the use of unavailable letters when interacting with the biased participant). Interestingly, this alignment extended to the group level, with five out of seven heterogeneous groups exhibiting group-level adaptation whereby the use of the biased letters significantly decreased over time in participants' languages (even when interacting with unbiased participants) (β=0.46±0.12, $p<0.001$) . Notably, the extent of this group-level adaptation was linked to the participants' initial learning accuracy, suggesting that stronger attachment to conventionalized forms can result in less accommodation to minorities. Together, our results show that individual differences in language use can spread to the wider community, and also accumulate over time, ultimately contributing to language changes.

Acknowledgments

We are grateful to the Van Gogh program of the Institut Français and the Dutch organization for internationalization in education (Nuffic) for their funding support.

References

Blasi, D. E., Moran, S., Moisik, S. R., Widmer, P., Dediu, D., & Bickel, B. (2019). Human sound systems are shaped by post-Neolithic changes in bite configuration. *Science*, *363*(6432). https://doi.org/10.1126/science.aav3218

Butcher, A. R. (2018). The special nature of Australian phonologies : Why auditory constraints on human language sound systems are not universal. *Proceedings of Meetings on Acoustics*, *35*(1), 060004. https://doi.org/10.1121/2.0001004

Dediu, D., Janssen, R., & Moisik, S. R. (2017). Language is not isolated from its wider environment : Vocal tract influences on the evolution of speech and language. *Language & Communication*, *54*, 9-20. https://doi.org/10.1016/j.langcom.2016.10.002

Dediu, D., Janssen, R., & Moisik, S. R. (2019). Weak biases emerging from vocal tract anatomy shape the repeated transmission of vowels. *Nature Human Behaviour, 3,* 1107–1115. https://doi.org/10.1038/s41562-019-0663-x

Fehér, O., Ritt, N., & Smith, K. (2019). Asymmetric accommodation during interaction leads to the regularisation of linguistic variants. *Journal of Memory and Language*, *109*, 104036. https://doi.org/10.1016/j.jml.2019.104036

Fehér, O., Wonnacott, E., & Smith, K. (2016). Structural priming in artificial languages and the regularisation of unpredictable variation. *Journal of Memory and Language*, *91*, 158-180. https://doi.org/10.1016/j.jml.2016.06.002

Jameson, K. A., & Komarova, N. L. (2009a). Evolutionary models of color categorization. I. Population categorization systems based on normal and dichromat observers. *J Opt Soc Am A Opt Image Sci Vis*, *26*(6), 6.

Jameson, K. A., & Komarova, N. L. (2009b). Evolutionary models of color categorization. II. Realistic observer models and population heterogeneity. *J Opt Soc Am A Opt Image Sci Vis*, *26*(6), 6.

Josserand, M., Allassonnière-Tang, M., Pellegrino, F., & Dediu, D. (2021). Interindividual Variation Refuses to Go Away : A Bayesian Computer Model of Language Change in Communicative Networks. *Frontiers in Psychology, 12,* 626118. https://doi.org/10.3389/fpsyg.2021.626118

Labov, W. (1972). *Sociolinguic Patterns* (Blackwell).

Labov, W. (1990). The intersection of sex and social class in the course of linguistic change. *Language Variation and Change, 2*(2), 205-254. https://doi.org/10.1017/S0954394500000338

Navarro, D. J., Perfors, A., Kary, A., Brown, S. D., & Donkin, C. (2018). When Extremists Win : Cultural Transmission Via Iterated Learning When Populations Are Heterogeneous. *Cognitive Science*, *42*(7), 7. https://doi.org/10.1111/cogs.12667

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

Raviv, L., Meyer, A., & Lev-Ari, S. (2019b). 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

Rita, M., Strub, F., Grill, J.-B., Pietquin, O., & Dupoux, E. (2022). *On the role of population heterogeneity in emergent communication*.

Tamminga, M. (2021). Macro and micro perspectives : Chapter 12. Leaders of language change. In H. Van de Velde, N. H. Hilton, & R. Knooihuizen (Éds.), *Language Variation – European Perspectives VIII: Selected papers from the Tenth International Conference on Language Variation in Europe (ICLaVE 10), Leeuwarden, June 2019* (p. 269-290). John Benjamins Publishing Company. https://doi.org/10.1075/silv.25.12tam

Yu, A. C. L., & Zellou, G. (2019). Individual Differences in Language Processing : Phonology. *Annual Review of Linguistics*, *5*(1), 131-150. https://doi.org/10.1146/annurev-linguistics-011516-033815

Who benefits from redundancy in learning noun class systems?

Aislinn Keogh*[1] and Gary Lupyan[2]

*Corresponding Author: aislinn.keogh@ed.ac.uk
[1]Centre for Language Evolution, University of Edinburgh, Edinburgh, UK
[2]Department of Psychology, University of Wisconsin-Madison, Madison, WI, USA

Redundancy is ubiquitous in the world's languages, but its functions are not yet well understood. Here, we propose that redundancy might contribute to the robustness of language to facilitate its learning by users with diverse cognitive traits. We use an artificial language learning experiment to identify individual differences in learning of noun classes from redundant linguistic cues. All logically possible behaviors are represented in our data: some participants prefer Cue A, some prefer Cue B, and some form a more holistic representation of Cue A+B. Despite this diversity, the population as a whole was above-chance when generalizing to novel stimuli, suggesting that redundancy helps people converge on similar surface structures, even if their underlying representations differ.

1. Introduction

All languages have a substantial degree of redundancy: the same information is encoded in multiple parts of the signal. For example, morphosyntactic elements such as agreement systems often involve marking words for features (e.g. person, gender, number or case) that are predictable from other cues (Haig & Forker, 2018). The pervasiveness of redundancy in language is a puzzle, especially in the face of evidence that producers prefer to minimize redundancy by omitting or reducing more predictable elements (e.g. Gibson et al., 2019; Jaeger, 2010; Aylett & Turk, 2004). One proposed explanation is that redundancy is a design feature that improves language learning, especially for young children (Tal & Arnon, 2022; Lupyan & Dale, 2010; Gerken et al., 2005; Morgan et al., 1987; Portelance et al., 2023) or the real-time processing of language (Christiansen & Chater, 2016). Another (non-mutually exclusive) possibility is that redundancy contributes to the *robustness* of language in the face of having to be acquired by diverse learners (Monaghan, 2017; Winter, 2014; Whitacre, 2010). Although there are cultural selection pressures against language structures that are not learnable by a large proportion of the population (Kirby et al., 2015), language systems may not be able to be optimized to be equally learnable by *all* members of a community, given the diversity in people's cognitive traits. Redundancy may be one way to increase the likelihood that a language will be learned equally well by everyone: even if some people fail to learn certain cues, they should still be able to learn the linguistic system overall.

Here, we offer an exploratory analysis of individual differences in learning of a redundant system. We present an experiment in which participants are trained on an artificial language with noun classes marked by redundant linguistic cues: a suffix on the noun, and a separate class marker. We then test how well they have learned these cues by asking them to generalize to novel meanings. Naturally, we expect to see variability in how well people learn the training set. However, our key interest is whether there is also variability in *generalization*, even among participants who appear to be performing similarly in training. A range of behaviors are logically possible: participants could learn the training set by rote without identifying any underlying rules or structure, they could learn both cues to class membership equally well, or they could learn the two cues to differing extents. We consider participants' training profiles and cognitive dispositions to try and predict who is more likely to exhibit these different behaviors.

2. Method

Participants We recruited 100 adults via Prolific. All resided in the US and were self-reported native English speakers with no known language disorders. Participants were paid $7.30 for around 45 minutes' participation.

Materials Stimuli were drawings from the MultiPic databank (Duñabeitia et al., 2018). We selected eight basic-level categories from four semantic domains: humans, animals, food, and clothing. The lexicon consisted of 32 pseudoword roots, four suffixes and four class markers taken from Culbertson et al. (2017). A full phrase consisted of a pre-nominal class marker followed by root + suffix e.g. *gae skun-po*. Phrases were displayed both auditorily and orthographically.

Procedure The experiment was written in JavaScript using the jsPsych library (de Leeuw, 2015) and administered through participants' web browser. First, participants were trained on a subset of the artificial language: four randomly selected meanings from each class. On each trial, participants heard a phrase and attempted to select its meaning from a 2x2 array of images: two from the target class, and two from another randomly selected class. They received full feedback on their selection. Participants completed 8 blocks of training, with each of the 16 meanings appearing as the target on one trial per block (128 trials total). Next, participants completed a reading span task to provide a measure of verbal working memory (Daneman & Carpenter, 1980; Friedman & Miyake, 2005), and a questionnaire assessing approach and avoidant behavioral tendencies (the BIS/BAS measurement tool: Carver & White, 1994). Both of these variables have been found to correlate with generalization performance in other domains (e.g. Dale et al., 2021). Participants were then tested on their knowledge of the language's structure using the held-out meanings. On each trial, participants heard an unfamiliar phrase and attempted to select its meaning from a 2x2 array of images: the target, and one randomly selected meaning from each of the other classes. There were three trial types in this phase. On REDUNDANT trials participants saw complete phrases as in training; on CLASSIFIER-ONLY and NOUN-ONLY trials they saw only one cue (the missing word was blanked out). They received no feedback on their selections. Finally, participants completed a questionnaire assessing explicit awareness of the noun classes and other language learning experience.

3. Results

Training Overall, participants showed clear evidence of learning over the course of training, with accuracy increasing considerably from the first block ($M = 0.43$, $SD = 0.49$) to the final block ($M = 0.84$, $SD = 0.37$). However, there are noticeable differences between participants. Visual inspection of by-participant loess curves reveals that there are at least three qualitatively different training profiles (Fig. 1A): linear (accuracy continues to increase throughout the training phase), logistic (accuracy increases from the start of training but ultimately reaches an asymptote) and non-monotonic (accuracy varies across the training phase). Controlling for explicit awareness of the semantic categories, higher performance in the final training block was predicted by higher performance in the first block ($\beta = 0.041$, $SE = 0.015$, $t = 2.72$, $p < 0.01$), higher working memory capacity ($\beta = 0.036$, $SE = 0.016$, $t = 2.27$, $p < 0.05$) and greater experience with language-learning apps like Duolingo ($\beta = 0.032$, $SE = 0.015$, $t = 2.14$, $p < 0.05$).

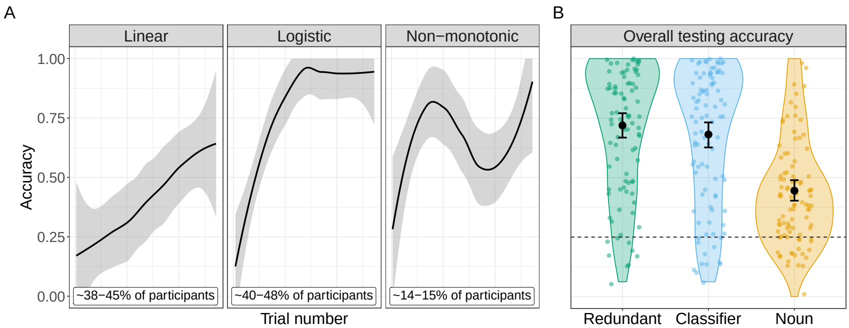


Figure 1. **A:** Example training profiles from three characteristic participants. **B:** Accuracy on test trials by cue. Individual coloured points represent by-participant mean performance for that cue. Black points and error bars represent the mean and bootstrapped 95% confidence interval over participants. The dashed line indicates chance performance of 0.25.

Generalization Because participants had not previously been exposed to the specific meanings presented at test, they could not rely on knowledge of the roots to determine the meaning of the phrases. Rather, the only way they could succeed at this task was if they had learned the relationship between the classifier and/or suffix and the semantic class. A larger drop in accuracy between the final training block and the test phase indicates that a person is less good at learning that relationship. Importantly, someone can be very good at memorizing the specific phrase-item pairings presented in training (high training accuracy) and yet fail to generalize, resulting in low accuracy on the test trials.

Overall, performance was above chance for all trial types, indicating that at the population-level, there is generalization of class cues (Fig. 1B). Accuracy was highest for REDUNDANT trials (M = 0.72, SD = 0.45), closely followed by CLASSIFIER-ONLY trials (M = 0.68, SD = 0.47). NOUN-ONLY trials had considerably lower accuracy (M = 0.45, SD = 0.50). Unsurprisingly, accuracy on the final training block predicted overall test performance (t = 3.47, $p < 0.001$). Surprisingly, this relationship (r = 0.11) all but disappeared when we include measures of working memory (β = 0.075, SE = 0.021, t = 3.530, $p < 0.001$), risk aversion (β = 0.043, SE = 0.019, t = 2.31, $p < 0.05$) and explicit awareness of the association between word forms and semantic categories (β = 0.099, SE = 0.040, t = 2.45, $p <$ 0.05). These three covariates all independently predicted test performance while controlling for final training block accuracy (β = 0.036, SE = 0.031, t = 1.796, p = 0.076), together accounting for 32% of the variance. Thus, better learners were not necessarily better generalizers.

Unsurprisingly, there was a clear drop-off in accuracy from the final training block to test (averaging across trial types: M = -0.22, SD = 0.23). A larger drop (controlling for training performance) is consistent with people being more focused on memorizing the specific items than on learning the underlying rules. Looking just at REDUNDANT test trials (the most like-for-like comparison), higher working memory capacity was associated with a smaller drop-off (β = -0.11, SE = 0.025, t = -4.23, $p < 0.001$). Higher reward responsiveness was associated with a slightly *larger* drop-off (β = 0.051, SE = 0.023, t = 2.21, $p < 0.05$), potentially due to the lack of feedback during testing (positive feedback may be viewed as a kind of reward).

Almost no one had uniform performance on the three test trial types, suggesting that the vast majority of participants had a preferred cue. We calculated two indices for every participant to compare their average performance on REDUNDANT trials to each of the individual cues. A larger positive score for the comparison between Cue A and REDUNDANT trials indicates that a participant is relying more on *Cue B*, since their performance is more greatly impaired by the removal of that cue. A variety of behaviors are represented in our data (Fig. 2), but participants clearly tend to rely more on the classifier than the suffix. Only 21 participants performed equally well or better on NOUN-ONLY trials relative to REDUNDANT trials, compared to 55 who performed equally well or better on CLASSIFIER-ONLY trials. Many (32) participants had positive scores on both indices, indicating that they were specifically benefiting from the redundancy i.e. had learned the association between classifier+suffix and semantic category in a more holistic way such that their performance declined when either cue was missing. Controlling for raw performance on REDUNDANT trials, higher performance in the final training block was associated with an increased benefit of redundancy (β = 0.047, *SE* = 0.011, *t* = 2.96, $p < 0.01$), while higher performance in the first half of the training phase (i.e. faster learning) was associated with a reduced benefit of redundancy (β = -0.043, *SE* = 0.012, *t* = -3.15, $p < 0.001$). In fact, faster learners actually performed *worse* when presented with redundant cues compared to their preferred cue in isolation. Greater risk aversion was also associated with a lower redundancy advantage (β = -0.028, *SE* = 0.010, *t* = -2.88, $p < 0.01$).

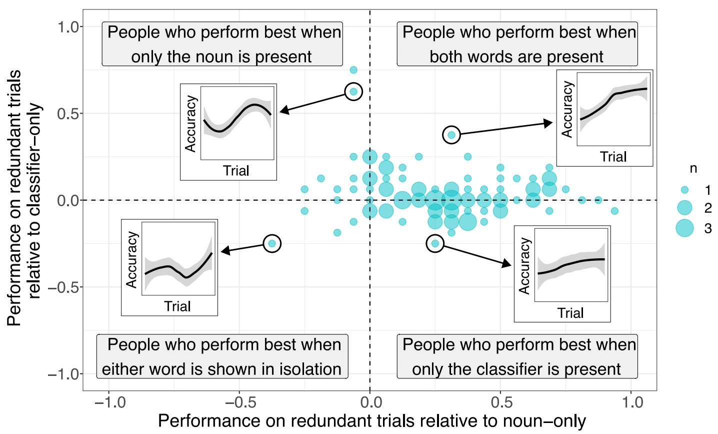


Figure 2. Performance on REDUNDANT trials relative to individual cues. Individual points represent by-participant scores; larger points represent more participants with equivalent values. Positive scores indicate a facilitatory effect of redundancy; negative scores indicate a detrimental effect of redundancy. Points along the dashed lines indicate that performance is equally good on REDUNDANT trials as on the given individual cue. Insets show the learning curves for the highlighted participants.

4. Discussion

In this study, we investigated whether morphosyntactic redundancy could contribute to the robustness of language by providing greater assurance that a system will be acquired despite variability in learning mechanisms across a population. When trained on an artificial language with two linguistic cues to noun class membership, we found clear individual differences in cue preference. Although the majority of our participants relied more heavily on the separate class-marker, a sizeable minority were attending more to the suffix on the noun. Around a third of our participants showed evidence of integrating the two cues more closely, performing best when both cues were available. This redundancy benefit was greatest for participants who achieved the highest level of accuracy by the end of the training phase, and lowest for participants who reached a higher level of accuracy earlier on in training.

The lack of a redundancy advantage for faster learners suggests that early commitment to one cue that reliably predicts category membership may block discovery of additional generalizations that might be beneficial down the line (in classical conditioning terms, *overshadowing*: Pavlov, 1927). Learners who explore the data for longer may be better able to integrate the redundant cues, and use these extra sources of information to their advantage both in learning and in generalization (Liquin & Gopnik, 2022; Sumner et al., 2019). Higher risk aversion also appears to reduce the strength of this overshadowing effect, resulting in more even performance across the three trial types, and therefore a lower benefit of redundancy *per se*. It is important to note that this is not a straightforward consequence of these participants expending greater effort: a person could be trying very hard during training, yet fail to learn the structure in a way that enables them to generalize training data effectively.

Contrary to some previous work in the 'Less-is-More' tradition (e.g. Goldowsky & Newport, 1993; Kareev, 1995; Pitts Cochran et al., 1999), we also found a positive relationship between working memory capacity and generalization. This finding dovetails with more recent work arguing that enhanced cognitive capacity is associated with better L1 and L2 learning outcomes (e.g. Brooks & Kempe, 2019; Rohde & Plaut, 2003), as well as studies linking higher working memory capacity to better category learning (e.g. Craig & Lewandowsky, 2012).

Our study also offers preliminary evidence of robustness effects in morphosyntax. Future work can implement different training conditions to see whether, at a population-level, there is better generalization of a language with redundant cues than one with a single cue – even if that single cue is well-learned by the majority of participants, like the classifier in our experiment. Manipulating the reliability of the redundant cues (Monaghan et al., 2017) may also force people to attend to both cues, reducing individual differences in cue preference. In fact, it is possible that even those participants who seemed to benefit from redundancy did not interpret the cues as redundant *per se*: since both were always available in training, they could have been interpreted as a single discontinuous cue.

Overall, this study adds to a growing body of evidence suggesting that, despite its potential costs in production, redundancy may be functional for language learning. Specifically, we suggest that when multiple cues to a language's grammatical structure are available, learners who favour different cues should nonetheless be able to acquire that underlying structure equally well.

Acknowledgements

This project received funding from the Economic and Social Research Council (ref. ES/P000681/1, held by AK) and NSF-PAC (ref. 2020969, held by GL). We are grateful to Kira Breeden for recording the audio stimuli.

References

Aylett, M., & Turk, A. (2004). The Smooth Signal Redundancy Hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, *47*(1), 31–56.

Brooks, P. J., & Kempe, V. (2019). More is more in language learning: Reconsidering the Less-Is-More Hypothesis. *Language Learning*, *69*, 13–41.

Carver, C. S., & White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS Scales. *Journal of Personality and Social Psychology*, *67*(2), 319–333.

Christiansen, M. H., & Chater, N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*.

Craig, S., & Lewandowsky, S. (2012). Whichever way you choose to categorize, working memory helps you learn. *Quarterly Journal of Experimental Psychology*, *65*(3), 439-464.

Culbertson, J., Gagliardi, A., & Smith, K. (2017). Competition between phonological and semantic cues in noun class learning. *Journal of Memory and Language*, *92*, 343–358.

Dale, G., Cochrane, A., & Green, C. S. (2021). Individual difference predictors of learning and generalization in perceptual learning. *Attention, Perception, & Psychophysics*, *83*(5), 2241–2255.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*(4), 450–466.

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*(1), 1–12.

Duñabeitia, J. A., Crepaldi, D., Meyer, A. S., New, B., Pliatsikas, C., Smolka, E., & Brysbaert, M. (2018). MultiPic: A standardized set of 750 drawings with norms for six European languages. *Quarterly Journal of Experimental Psychology (2006)*, *71*(4), 808–816.

Friedman, N. P., & Miyake, A. (2005). Comparison of four scoring methods for the reading span test. *Behavior Research Methods*, *37*(4), 581–590.

Gerken, L., Wilson, R., & Lewis, W. (2005). Infants can use distributional cues to form syntactic categories. *Journal of Child Language*, *32*(2), 249–268.

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*, *23*(5), 389–407.

Goldowsky, B. N., & Newport, E. L. (1993). Modeling the effects of processing limitations on the acquisition of morphology: The less is more hypothesis. *The proceedings of the 24th Annual Child Language Research Forum*(February), 124–138.

Haig, G., & Forker, D. (2018). Agreement in grammar and discourse: A research overview. *Linguistics*, *56*(4), 715–734.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*(1), 23–62.

Kareev, Y. (1995). Through a narrow window: working memory capacity and the detection of covariation. *Cognition*, *56*(3), 263–269.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102.

Liquin, E. G., & Gopnik, A. (2022). Children are more exploratory and learn more than adults in an approach-avoid task. *Cognition*, *218*, 104940.

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PLOS ONE*, *5*(1), e8559.

Monaghan, P. (2017). Canalization of language structure from environmental constraints: A computational model of word learning from multiple cues. *Topics in Cognitive Science*, *9*(1), 21–34.

Monaghan, P., Brand, J., Frost, R., & Taylor, G. (2017). Multiple variable cues in the environment promote accurate and robust word learning. In *Proceedings of the 39th Cognitive Science Society Conference.*

Morgan, J. L., Meier, R. P., & Newport, E. L. (1987). Structural packaging in the input to language learning: Contributions of prosodic and morphological marking of phrases to the acquisition of language. *Cognitive Psychology*, *19*(4), 498-550.

Pavlov, P. I. (1927). *Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex.* London: Oxford University Press.

Pitts Cochran, B., McDonald, J. L., & Parault, S. J. (1999). Too smart for their own good: The disadvantage of a superior processing capacity for adult language learners. *Journal of Memory and Language*, *41*(1), 30–58.

Portelance, E., Duan, Y., Frank, M. C., & Lupyan, G. (2023). Predicting age of acquisition for children's early vocabulary in five languages using language model surprisal. *Cognitive Science*, *47*(9), e13334.

Rohde, D. L. T., & Plaut, D. C. (2003). Less is less in language acquisition. In P. Quinlin (Ed.), *Connectionist Modelling of Cognitive Development.* Hove, UK: Psychology Press.

Sumner, E., Li, A., Perfors, A., Hayes, B., Navarro, D., & Sarnecka, B. (2019). The exploration advantage: Children's instinct to explore allows them to find information that adults miss. *PsyArXiv Preprints*.

Tal, S., & Arnon, I. (2022). Redundancy can benefit learning: Evidence from word order and case marking. *Cognition*, *224*, 105055.

Whitacre, J. M. (2010). Degeneracy: a link between evolvability, robustness and complexity in biological systems. *Theoretical Biology and Medical Modelling*, *7*(1), 6.

Winter, B. (2014). Spoken language achieves robustness and evolvability by exploiting degeneracy and neutrality. *BioEssays*, *36*(10), 960–967.

Uncommon sounds for common words: Balancing phonemic type and token frequency within words for a higher entropy lexicon

Adam King*[1], Andrew Wedel[2]

*Corresponding Author: adam.king.phd@gmail.com
[1]GumGum, Santa Monica, USA
[2]Department of Linguistics, University of Arizona, Tucson, USA

It is a common empirical finding that the token frequency of a phonological contrast is positively correlated with its type frequency; sounds that occur most frequently across a corpus tend to appear in more unique word types and vice versa. This would naturally follow if contrasts are randomly distributed across words, though, given that sounds are used to distinguish words, this is precisely the opposite of what we expect. Lexical access proceeds incrementally, such that each successive sound gives a listener the required information to exclude a growing set of incompatible words as the speech stream is perceived (van Son & Pols 2003; Magnuson et al. 2007). If language is shaped to be an efficient system of communication, it would be expected that words in the lexicon would share similarities with a Huffman code, where each contrast equally divides the remaining words into probabilistically equal groups, causing instead a negative relation between the token frequency of a sound and the number of words it appears in at each branch-point, creating a more balanced contrast (Fig 1).

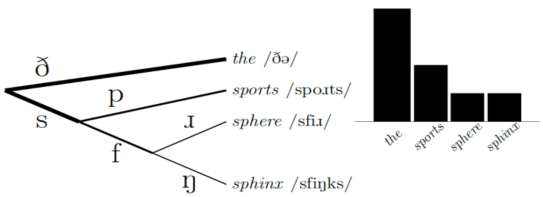


Figure 1. Contrast structure of a toy version of English. The bar chart on the right represents the probability of each word. On the left is a branching structure diagram of word contrasts where line thickness represents summed word probabilities along that branch. Each successive branch divides the remaining potential words into probabilistically balanced groups.

Here, we provide evidence for this predicted inverse relationship in a genetically balanced set of 20 languages, comparing the type and token frequencies of word-initial biphones. First, we show that the commonly-noted positive type/token correlation is an artifact of the floor formed by the least frequent words in a corpus: if a word occurs only once, the sounds in it must occur at least once; if a word occurs twice, the sounds within it must appear twice, etc. When biphones that make up this floor are removed, that is, word-initial biphones that together comprise less than 5%[1] of all tokens in the corpus, we find instead a strong *negative* correlation between biphone type frequency and the mean token frequency of the words in which the biphone is found (Fig 2), leading to a more balanced contrast than would be expected if efficient phonological information transmission were not a shaping force on the lexicon.

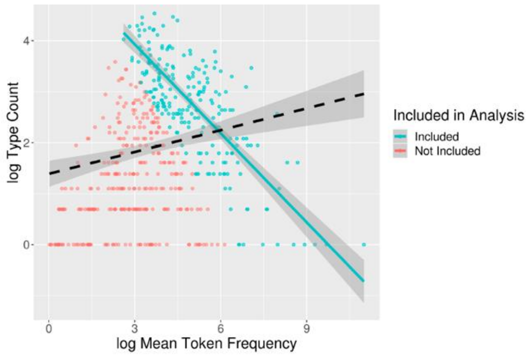


Figure 2. Relationship between mean token frequency and type count for word-initial biphones, excluding the least frequent 5% in the lexicon. For the top 95%, sounds that appear in fewer words tend to appear in higher frequency words overall, creating a more balanced contrast.

How might this balance in lexicons arise throughout language change? A large body of evidence shows that high-information segments tend to be hyperarticulated, while lower-information segments tend to be reduced (e.g., van Son & Pols 2003, Wedel et al. 2018), causing less frequent words to retain complex, marked segments over time, while more frequent words do the opposite. As less frequent words make up the large majority of a language's unique word types, these patterns of change should result in a lexicon in which less frequent words are composed of a wider diversity of sounds and sound sequences, as found in King & Wedel (2020). As a result, at a contrast point, sounds that are more often found in less frequent words should lead to relatively larger remaining cohorts, resulting in the expected negative type/token correlation. This represents a plausible mechanistic pathway from speaker-level micro-effects of information transmission toward a lexicon that is structured for higher entropy.

[1] The pattern remains the same for different threshold, e.g., 1%, 10%. We use 5% here for visualization purposes.

References

King, A., & Wedel, A. (2020). Greater early disambiguating information for less-probable words: The lexicon is shaped by incremental processing. *Open Mind*, *4*, 1-12.

Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K. & Aslin, R. N. (2007) The dynamics of lexical competition during spoken word recognition. *Cognitive Science 31*, 133-156.

Van Son, R. J. J. H. & Pols, Louis C. W. (2003). How efficient is speech. *Proceedings of the institute of phonetic sciences, 25, 171-184. University of Amsterdam*.

Wedel, A., Nelson, N., & Sharp, R. (2018). The phonetic specificity of contrastive hyperarticulation in natural speech. *Journal of Memory and Language, 100,* 61-88.

Communication and Linguistic Structure in Collaborative Human-Machine Language Evolution

Tom Kouwenhoven[*1], Neval Kara[2], and Tessa Verhoef[1]

[*]Corresponding Author: t.kouwenhoven@liacs.leidenuniv.nl
[1]Creative Intelligence Lab, Leiden University, Leiden, The Netherlands
[2]Department of Psychology, Çankaya University, Ankara, Turkey

The compositionality of human language allows us to combine meaningful words into more complex meanings. The emergence of compositionality is extensively examined through human experiments (e.g., Kirby et al., 2008, 2015; Raviv et al., 2019) and (agent-based) simulations (e.g., Kirby, 1998; Brighton, 2002; Vogt, 2005; Lazaridou & Baroni, 2020). The latter is seeing increased attention due to computational advances and a rising interest into large language models (LLM). Although the behaviour of LLMs is fundamentally different from humans, their linguistic abilities are unprecedented, rendering them the first close comparators of language users. Moreover, LLMs are capable of in-context learning, i.e., having the model tackle a novel task based on a few examples in the prompt (Brown et al., 2020). In this pilot study, we examine whether LLMs can act as controlled variables in experiments on language evolution. Specifically, we assess if linguistic structure evolves when participants ($n = 10$) communicate with an LLM (*text-davinci-003*, temperature 0.0) in the Lewis signalling game.

The setup of our pilot is based on that of Kirby et al. (2008) and Raviv et al. (2019). Participants go through an exposure phase to learn an initially holistic artificial language, followed by a labelling phase in which they type labels for each object. The LLM learns artificial languages through the in-context learning method used by Galke et al. (2023), who showed that LLMs can learn artificial languages and that, similar to experimental findings, systematic generalisation was higher for more structured languages. Following the labelling phase, the participants and LLM alternate roles (speaker, listener) and use the learned language to communicate for six rounds in a referential task with four objects and one target. Here, the LLM must generalise labels for unseen stimuli based on its context, i.e., the vocabulary. Similar to Raviv et al. (2019), the meaning space expands by three objects after each communication round, starting with 12 objects. Finally, the participant labels each scene in the naming phase. Accuracy is the percentage of correctly identified objects. Identical to Kirby et al. (2008), z-scores of the Mantel test (Mantel, 1967) indicate the degree of structure in the vocabulary.

Results

Although learning the initial holistic language proves difficult, participants can identify objects based on the labels to some extent ($\approx 60\%$ accuracy) after the exposure phase. However, using this language for communication is challenging (figure 1, left) since accuracy is not better than chance and does not improve over the rounds. Across the four phases of the experiment, we observe an increase in structure in the labels produced by human participants. Yet, the overall structure does not increase as radically as previously found in dyadic interactions, even though in principle the LLM can generalise over such artificial languages (Galke et al., 2023). We expect this is due to the computational mechanisms used to generate responses. While humans update their beliefs following interactions, LLMs only have access to the prompt containing the vocabulary and question at hand but do not integrate past experiences (e.g., (un)successful rounds) or reason about the others' state of mind. Our preliminary findings suggest that linguistic structure does not radically change over time, thereby not following findings from earlier experiments (Kirby et al., 2008; Raviv et al., 2019) with humans alone.

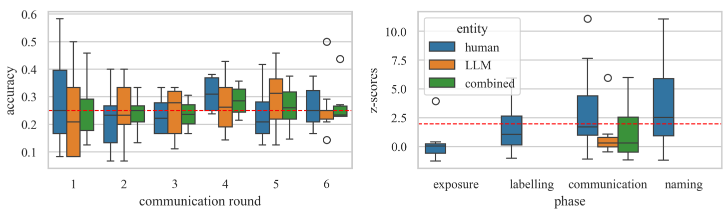


Figure 1. Accuracy over communication rounds (left) and the degree of structure (right) in each phase (only round 6 for the communication phase). The colour indicates whether predictions are from the human (blue), the LLM (orange), or both (green) entities. The red line indicates *chance* performance (left) and the threshold for which structure is likely to be caused by the entities instead of chance (right). Structure scores can be compared between exposure-labelling and communication-naming.

These results show that, although communication is difficult, the vocabularies do not completely collapse when used in human-machine communication. This is promising given that experiments with human-human dyads have shown that interaction dynamics and personal differences affect how language evolves (Verhoef et al., 2022; Kouwenhoven et al., 2022). Moreover, it may suggest that mere presence of a communicative partner–even one that is bad–can push the participant to increase the systematicity of its productions. While this work is preliminary and only a first step in human-machine language evolution, it opens possibilities for future work in which one communicative partner can be relatively fixed and kept under experimental control. For example, in more complex setups like iterated learning or by comparing languages from human-human with human-computer experiments to isolate biases of a single human in cooperative settings.

References

Brighton, H. (2002). Compositional syntax from cultural transmission. *Artificial life*, *8*(1), 25–54.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 1877–1901). Curran Associates, Inc.

Galke, L., Ram, Y., & Raviv, L. (2023). What makes a language easy to deep-learn? *arXiv preprint arXiv:2302.12239*.

Kirby, S. (1998). Syntax without natural selection: How compositionality emerges from vocabulary in a population of learners.

Kirby, S., Cornish, H., & 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., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102.

Kouwenhoven, T., Kleijn, R. de, Raaijmakers, S., & Verhoef, T. (2022). Need for structure and the emergence of communication. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44).

Lazaridou, A., & Baroni, M. (2020). Emergent multi-agent communication in the deep learning era. *arXiv preprint arXiv:2006.02419*.

Mantel, N. (1967). The Detection of Disease Clustering and a Generalized Regression Approach. *Cancer Research*, *27*, 209-220.

Raviv, L., Meyer, A., & Lev-Ari, S. (2019). Compositional structure can emerge without generational transmission. *Cognition*, *182*, 151-164.

Verhoef, T., Walker, E., & Marghetis, T. (2022). Interaction dynamics affect the emergence of compositional structure in cultural transmission of space-time mappings. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44).

Vogt, P. (2005). The emergence of compositional structures in perceptually grounded language games. *Artificial Intelligence*, *167*(1), 206-242. (Connecting Language to the World)

Pointing prevents the emergence of symbolic referential signals: empirical tests in a common task framework

Kateryna Krykoniuk*[1], Fiona M. Jordan[2] and Seán G. Roberts[3]

*Corresponding Author: k.krykoniuk@sheffield.ac.uk
[1]Department of Russian and Slavonic Studies, University of Sheffield, Sheffield, UK
[2]Anthropology and Archaeology, University of Bristol, Bristol, UK
[3]School of English, Communication and Philosophy, Cardiff University, Cardiff, UK

Two experiments tested the conditions under which symbolic referential communication systems emerge. Naive participants were placed in an 'arena' with conditions that were predicted to motivate the emergence of a symbolic system. However, this failed to occur in an arena based on collaborative building under several conditions due to the effectiveness of pointing. In contrast, an arena based on maintaining a fire provided a need to communicate distal meanings and led to greater likelihood of symbol system emergence. We present a tentative causal model that explains the results and suggests future studies.

1. Introduction

Pointing has been invoked as an important tool in the development of symbolic referential signals (SRS): a system of shared, symbolic signals that refer to objects in the world. Pointing can help ground the meaning of a signal by establishing joint attention between two interlocutors on a referent (e.g. Steels & Belpeame, 2005). However, pointing is also a powerful communication tool in its own right. This study investigates what specific ecological or social conditions are required for a symbolic system to be effective over and above pointing. Borrowing from Hurford, we call these conditions the 'arena of language evolution' (Hurford, 1989; 1990). Roberts, Irvine & Jordan (2022) suggested that this can be investigated using a 'common task framework': a series of comparable practical simulations that forces researchers to specify their principles and test them against each other in controlled conditions. That is, a sound theory should be able to define an 'arena' including an environment, and a task for agents to complete that reflects relevant and plausible analogues of early hominid life. The agents should start exhibiting the predicted communicative behaviour under the right arena conditions. Previous studies found that, contrary to some theoretical predictions, SRS were unlikely to emerge in an arena where agents had to collaboratively build a structure out of coloured blocks (Irvine & Roberts, 2016). While participants had the ability to invent a symbolic convention to refer to different block types, they simply used pointing with trial-and-error. The suggestion was that the cost of setting up a symbolic system was too high in comparison to the effectiveness of pointing. However, perhaps the critical barrier was not pointing *per se*, but the low cost of the trial-and-error strategy. Another possibility is that the system of meanings was too simple to require a dedicated symbolic system. In section 2, we replicate and extend the previous experiment to test these possibilities by increasing the cost of destroying blocks (reducing the effectiveness of trial-and-error) and by increasing the number of block colours. In contrast, section 3 presents a new arena based on theories of fire maintenance.

2. Arena A: Building (or collaborative manipulation of objects)

Replicating Irvine et al. (2016), the following arena was set up in *Minecraft*:
Environment: A flat field with markers showing the outline of a building.
Task: Two participants needed to follow a plan to build an abstract building from coloured blocks. Participants were not allowed to speak, but they could knock on the table or 'gesture' via their avatar's movements in the game. Participants were given up to 20 minutes to complete the task.
Asymmetry of information: Each participant had half of the plan of the building. The plan was asymmetric and unsystematic, meaning that participants had to communicate to each other the location and colours of blocks.
Division of labour: The plan included 4 colours of blocks, but each participant was only able to place two colours. This is analogous to individuals being specialised in the use of specific building materials. Participants were allowed to destroy blocks of any colour. There were two additional conditions. The second condition is identical, except blocks took twice as much time to destroy. The third condition included 8 colours of blocks (4 unique colours each) instead of 4.
33 pairs participated (11 in each condition, a given participant only took part in one condition). The experiment was recorded and then participants filled out a questionnaire about their communication strategy, then they were informally interviewed. The videos, interviews and participant questionnaires were analysed for various categories of communication strategies. A pair was considered to have established an SRS if both participants' questionnaires reported the same communicative convention for identifying the colour of their blocks, or if the interview revealed such a system. The form of the signal could be anything (knocking, jumping, spinning etc.). Other strategies were identified from the data rather than being assumed *a-priori*, and are described below.

2.1. Results

Every pair of participants succeeded in establishing a strategy to solve the task. The typical procedure was that each participant would start by building some portion of their own side of the building. Then, they would realise that they needed the help of their partner and seek their attention. Pairs built one side at a time, so one participant would take the role of ‘director’, indicating locations and colours, and the other taking the role of the ‘builder’ who placed the blocks.

Table 1 shows the communication strategies that emerged in each condition. Note that these are not mutually exclusive. In the building situation, the predominant strategy was to use pointing to identify locations and trial and error to identify block colour: a director would indicate a place for the builder to place a block, but destroy it if it was the wrong type. To assist this strategy, the majority of all pairs established conventions for signaling ‘correct’ or ‘incorrect’, most often analogies of the real-world convention of nodding and shaking the head. While this is a symbolic convention, it does not refer to objects in the game world. Therefore, we did not take this as evidence for the emergence of symbolic referential signals.

Beyond trial and error, the dominant secondary strategy varied by condition. Two strategies used a feature of *Minecraft* where players can see the colour of the blocks their partners are currently holding. In the initial condition (4 colours), the most frequent secondary strategy was for the director to switch the blocks they themselves were holding. This was a cue for the builder to change the block type they were holding. While this relies on an analogy, the signal’s meaning is “change your block” or just “incorrect”. It was highly contextual and could not be used to refer to a specific colour in a different context. That is, it is an extension of the ‘trial and error’ system that avoids needing to place and destroy a block.

Table 1. Strategies adopted in each condition of the building arena (dominant strategy in bold).

Condition	*Correct/ Incorrect*	*Director switches blocks*	*Builder switches blocks*	*Indexical system*	*SRS*
4 colours	91%	**55%**	27%	18%	9%
Hard blocks	73%	9%	**45%**	9%	9%
8 colours	82%	18%	27%	**64%**	9%

In the ‘hard blocks’ condition, the dominant strategy shifted to a similar system, but this time the builder would choose a block type and wait for confirmation from the director that it was the right colour (e.g. by nodding). This is essentially the same as pointing at a candidate object, and is a logical strategy to adopt when placing the wrong block colour is a more costly mistake.

In the 8 colour condition, the dominant strategy was an indexical. Participants pointed to existing blocks to indicate the colour or placed a set of ‘reference blocks’ to one side of the main building to have access to a full set of colours. This strategy relies on pointing alone. It is more efficient than trial and error with the expanded number of colours, though directors sometimes spent time encouraging builders to place reference blocks in order to point at them.

In general, the condition affected the dominant strategy that emerged (Fisher’s exact test of director/builder/indexical strategy frequency, $p = 0.038$), indicating that the task demands were sufficiently different to motivate different communication strategies. However, only one pair in each condition established an SRS. In one case in the 4 blocks condition, the pair established a system immediately before doing anything else. One participant knocked once holding a blue block, knocked twice holding a red block, then knocked once holding a blue block. This redundant repetition signalled an ostensive action. Their partner understood the idea and did the same with their blocks. This process took only 17 seconds and was the strategy expected in the previous study. So, while establishing a symbolic communication system is clearly possible, the arena does not provide enough motivation for this to emerge frequently. Indeed, in many cases, participants reported that they had considered establishing a system, but decided that it was not worth the effort. In one condition, the pairs even managed to complete much of the task without communicating directly: one participant placed blocks randomly and the other destroyed incorrectly placed blocks. In summary, in the building arena, the ability to point at objects makes an SRS redundant. This suggests that an arena that motivates the emergence of an SRS needs to involve meanings that cannot be pointed to.

3. Arena B: Fire maintenance

There are many possible arenas with distal meanings. However, given the preference for pointing solutions, symbolic referential signals may only emerge when the arena discourages pointing strategies. Since participants are free to move around the world, there are few realistic scenarios where individuals can be prevented from going together to the referents. Put another way, if you can point at something to request someone to give it to you, you can just pick it up yourself. So the key property is that the most efficient solution should involve the 'director' needing to be distant from the referents at the point when they are requesting them. That is, the 'distal' property of meanings is not necessarily inherent to the meaning or referent, but emerges from an interaction between where the referents and interlocutors are in context. Fire maintenance may provide inspiration for such an arena (Twomey, 2013). Since fire use preceded fire making, fires from natural sources would need to be constantly monitored and maintained. However, fuel would also need to be gathered. This inspired an arena where participants had to collect raw materials and 'smelt' them into refined materials:

Environment: A narrow strip of land between a lake and a sheer mountain. A furnace was placed at one end near a source of fuel, and a 'mine' was placed at the other end with a source of gold ore and green ore. In an alternative condition, the mine was placed near to the furnace.

Task: Smelt the ores into 'ingots' by adding ore and fuel to the furnace.

Asymmetry of information: The 'smelter' had a set of cards that indicated the order in which ingots should be produced. This was not observable to the 'miner'.

Division of labour: The smelter was taught to use the furnace and the miner was taught to obtain ore from the mine, though there was no rule against swapping.

The rest of the methods were identical to arena A. The expected optimal strategy was for the smelter to communicate the type of ore required to the miner, then the miner gets the ore while the smelter gets fuel, monitors the furnace, and produces the ingots when the miner returns. The arena was designed so that the time it took to get ore was roughly equal to the time to ‘maintain’ the furnace and produce an ingot. This meant that, if the participants were acting efficiently, the smelter would remain at the furnace and both participants would not be in the same place with both types of ore, avoiding an opportunity for pointing at a required object. The analogue in the real world might be needing to tell someone to collect a specific type of fuel while they kept a fire going.

3.1 Results

Table 2 shows the frequency of established SRS for both arenas. The fire arena motivated 36% of the pairs to establish an SRS, marginally more frequently than the building arena (Fisher’s p = 0.053). More tellingly, at least one participant in each pair attempted to establish an SRS in 86% of trials, compared with only 12% of trials in the building arena (p = 0.0001). Only two fire trials did not attempt to establish a symbolic system, and in one of them a smelter reported that they would have done but thought that they were not allowed to go to the mine. In fact, we had to exclude one trial from the data because a participant shouted out a symbolic referential strategy to their partner before the experimenter had finished announcing the rules of the task. In contrast, when the mine was near, attempted SRS were significantly rarer (p = 0.002) and similar to the Building arena.

Table 2. The frequency of established and attempted SRS in each arena.

Arena	*Established SRS*	*Attempted SRS*
Building	9%	12%
Fire Maintenance (distant mine)	36%	82%
Fire Maintenance (near mine)	9%	9%

The symbolic system was usually established by both participants going to the mine, pointing to a type of ore and producing a signal with a knock. That is, pointing was important to help ground the signals, but the necessity of signals was created by the conditions of the arena.

4. Discussion

The building arena consistently failed to motivate a symbolic referential communication system. Instead, participants found creative solutions involving pointing. Manipulating the opportunity cost and the number of meanings changed the secondary communication strategy, but did not affect the likelihood of referential symbols emerging. In contrast, SRS frequently emerged in the fire arena, and participants were more likely to report feeling a need for such a system. A tentative causal model is suggested in figure 1: Conditions of the arena create pressures for specific types of social interaction, and these change the effectiveness of various communication strategies. The core of the model is that asymmetry of information and division of labour create a need for communication, and this increases the effectiveness of a pointing or symbol system and the likelihood of one emerging. Asymmetry of information was provided by dividing the building plan in the building arena and by specialised training and only the smelter knowing the ore sequence in the fire arena. Division of labour was motivated by having specialised blocks in the building arena but was more complex in the fire arena: specialized training, a need to monitor the fire, and a distant mine. The effectiveness of an SRS is affected by the need for communication, how much time the symbol system costs to set up, and how effective a pointing strategy is. If pointing can solve the task, this directly prevents the need for a symbol system and indirectly increases the opportunity cost of setting one up. Increasing the number of blocks was predicted to reduce the effectiveness of trial-and-error, but it also increased the cost of setting up a symbol system. Increasing the hardness of blocks was predicted to increase the cost of mistakes, so reduce the relative cost of setting up a symbol system since both pointing and symbols would provide more confidence. However, apparently it did not reduce the relative effectiveness of the pointing system enough for a symbolic system to emerge. In contrast, the fire arena created distal meanings: the most efficient task solution involved the smelter needing to request an item when it was not immediately present. This reduced the effectiveness of pointing, reducing the relative cost of setting up a symbol system (compared to both participants travelling to the mine) and motivated participants to invent an SRS.

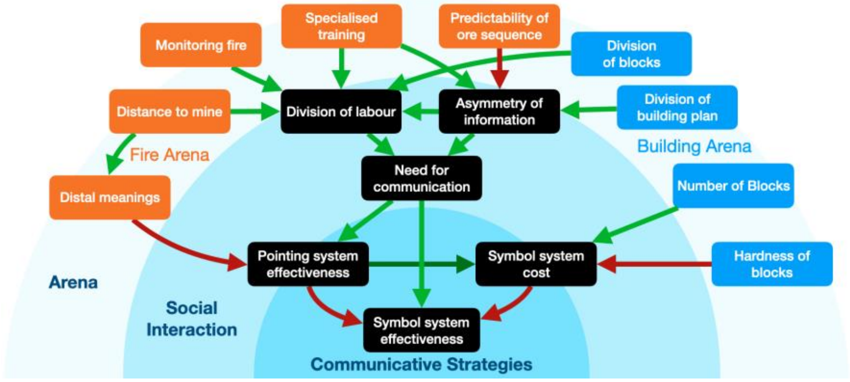


Figure 1. A causal model of symbol system emergence. Green: positive effect, red: negative effect.

There may be many other arenas that motivate the evolution of symbols, including negotiating the division of labour, teaching, or the need to refer to meanings that distant in time. The common task framework can also test and contrast these.

Acknowledgements

KK and SGR were supported by an AHRC grant AH/T006927/1.

References

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. (1990). Nativist and functional explanations in language acquisition. *Logical issues in language acquisition*, *85*, 136.

Irvine E. and Roberts S. (2016). Deictic Tools Can Limit The Emergence Of Referential Symbol Systems. In S.G. Roberts, C. Cuskley, L. McCrohon, L. Barceló-Coblijn, O. Fehér & T. Verhoef (eds.) The Evolution of Language: Proceedings of the 11th International Conference (EVOLANG11). Available online: http://evolang.org/neworleans/papers/99.html

Roberts, S. (2018). What are the social, economic and ecological conditions for the evolution of complex communication systems? Comment on "Rethinking foundations of language from a multidisciplinary perspective" by T. Gong et al. *Physics of Life Reviews* 26 (152-154).

Roberts, S., Irvine, E. & Jordan, F. (2022). The Common Task Framework: Using Causal Theories From Linguistic Anthropology to Explore the Emergence of Symbolic Communication. In Ravignani, A., Asano, R., Valente, D., Ferretti, F., Hartmann, S., Hayashi, M., Jadoul, Y., Martins, M., Oseki, Y., Rodrigues, E. D., Vasileva, O. & Wacewicz, S. (Eds.): Proceedings of the Joint Conference on Language Evolution (JCoLE). doi:10.17617/2.3398549.

Steels, L., & Belpaeme, T. (2005). Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and brain sciences*, *28*(4), 469-488.

Twomey, T. (2013). The cognitive implications of controlled fire use by early humans. *Cambridge Archaeological Journal*, *23*(1), 113-128.

Middle Pleistocene humans in Europe: cognition before Neanderthals

Svetlana Kuleshova[1,2]

*Corresponding Author: 40010189@parisnanterre.fr
[1]Center for Language Evolution Studies, Nicolaus Copernicus University in Toruń, Toruń, Poland
[2]ArScAn-Équipe AnTET (UMR 7041), CNRS, Université Paris Nanterre, Nanterre, France

Despite the ongoing debate about the complexity of the Neanderthal mind (Romagnoli, Rivals, and Benazzi, 2022), their cognitive development is now widely recognized in the scientific community. Neanderthals, initially perceived as primitive humans, were rehabilitated by the researchers and are now often considered cognitively very close to *Homo sapiens* (Otte, 2019; Slimak, 2019). On the basis of their complex technology, they are often supposed to have been capable of planning several steps ahead, thus displaying enhanced working memory capacity (Sykes, 2015). It is also hypothesized that Neanderthals also demonstrated some symbolic behavior by burying their dead and using pigments and ornaments (Dediu and Levinson, 2018). All these findings allowed researchers to suppose that Neanderthals may have possessed some kind of protolanguage, or even a recognizably modern language (Dediu and Levinson, 2018; Johansson, 2014, but see Berwick, Hauser, and Tattersall, 2013).

The situation is, however, different when it comes to the ancestors of Neanderthals. Indeed, the cohabitation of different human species is often referred to as "muddle in the middle." The discussions between paleoanthropologists about the number and origins of species in Europe in the early and mid-Middle Pleistocene are still ongoing (Dennell, Martinón-Torres, and Bermúdez de Castro, 2011; Athreya and Hopkins, 2021). *Homo heidelbergensis* is the most well-known early Middle Pleistocene species identified today in Europe (although the particular characteristics of the species are still debated), but arguably it was not the only one (de Lumley, 2015). In this poster, we present a comprehensive review of the existing evidence about the cognitive capacities of *Homo heidelbergensis*, who most probably was the direct ancestor of Neanderthals (Di Vincenzo and Manzi, 2023). However, because of ongoing debates and the frequent impossibility of associating archaeological material with fossil records, it is often problematic to assign a particular archaeological collection to a particular species and thus to distinguish between the material cultures of potentially different species. Here, *Homo heidelbergensis* is used to refer collectively to early and mid-Middle Pleistocene humans in Europe. By comparing this evidence with what is already known about Neanderthals, we can better understand what cognitive capacities were potentially inherited and which were likely developed in this new species.

The possibility of the existence of a protolanguage before Neanderthals is discussed through a thorough review of existing arguments. We consider the arguments based on anatomical evidence and on the material culture left by those populations. Anatomical evidence shows that *Homo heidelbergensis* was right-handed (Faurie, Raymond and Uomini, 2016) and had a vocal tract similar to the Neanderthal one (Martínez et al., 2013), which might indicate brain specialization and a potential for speech. On the material culture side, we find elaborated habitat structures (de Lumley, 2006), complex Acheulean technology executed on different raw materials, and regionally distinguishable traditions by the end of the period (Carrión and Walker, 2019; Davis and Ashton, 2019), as well as occupation of northern Europe (Hosfield and Cole, 2018). These findings suggest an enhanced ability for planning (Hosfield and Cole, 2018), cultural transmission, and social organization (Ashton and Davis, 2021). Stout et al. (2014) suggest hierarchical behaviour organization. The analysis of these arguments leads to a better understanding of the evolution of the cognition of early humans populating Europe during the Lower Palaeolithic. After reviewing the *Homo heidelbergensis* evidence, we present a comparative section to contrast the arguments and better understand the cognitive evolution between the two species. In light of this newly emerged evidence, it’s argued that the difference between the cognitive capacities of Neanderthals and their ancestors is rather a matter of degree, and to the extent that this is indicative of language (but see Bar-Yosef, 2017; Botha, 2011), to get to the origins of language, maybe we should take another step back in time and turn our attention to these earlier populations.

Acknowledgements

I thank Slawomir Wacewicz and Michael Pleyer for valuable comments.

References

Ashton, N., & Davis, R. (2021). Cultural mosaics, social structure, and identity: The Acheulean threshold in Europe. *Journal of Human Evolution*, *156*, 103011.

Athreya, S., & Hopkins, A. (2021). Conceptual issues in hominin taxonomy: Homo heidelbergensis and an ethnobiological reframing of species. *American Journal of Physical Anthropology*, *175*(S72), 4–26.

Bar-Yosef, O. (2017). Can archaeology tell us about the evolution of cognition and language? *Journal of Neurolinguistics*, *43*, 222–227.

Berwick, R., Hauser, M., & Tattersall, I. (2013). Neanderthal language? Just-so stories take center stage. *Frontiers in Psychology*, *4*.

Botha, R. (2011). Inferring Modern Language from Ancient Objects. In M. Tallerman & K. R. Gibson (Eds.), *The Oxford Handbook of Language Evolution* (pp. 303–312). Oxford: Oxford University Press.

Carrión, J. S., & Walker, M. J. (2019). Background to Neanderthal presence in Western Mediterranean Europe. *Quaternary Science Reviews*, *217*, 7–44.

Davis, R., & Ashton, N. (2019). Landscapes, environments and societies: The development of culture in Lower Palaeolithic Europe. *Journal of Anthropological Archaeology*, *56*, 101107.

de Lumley, H. (2006). Il y a 400 000 ans: La domestication du feu, un formidable moteur d'hominisation. *Comptes Rendus Palevol*, *5*(1), 149–154.

de Lumley, M.-A. (2015). L'homme de Tautavel. Un *Homo erectus* européen évolué. *Homo erectus tautavelensis*. *L'Anthropologie*, *119*(3), 303–348.

Dediu, D., & Levinson, S. C. (2018). Neanderthal language revisited: Not only us. *Current Opinion in Behavioral Sciences*, *21*, 49–55.

Dennell, R. W., Martinón-Torres, M., & Bermúdez de Castro, J. M. (2011). Hominin variability, climatic instability and population demography in Middle Pleistocene Europe. *Quaternary Science Reviews*, *30*(11), 1511–1524.

Di Vincenzo, F., & Manzi, G. (2023). Homo heidelbergensis as the Middle Pleistocene common ancestor of Denisovans, Neanderthals and modern humans. *Journal of Mediterranean Earth Sciences*, *15*.

Faurie, C., Raymond, M., Uomini, N. (2016). Origins, Development, and Persistence of Laterality in Humans. In F. Loffing, N. Hagemann, B. Strauss, C. MacMahon (Eds.), Laterality in Sports (pp. 11-30). Amsterdam: Academic Press.

Hosfield, R., & Cole, J. (2018). Early hominins in north-west Europe: A punctuated long chronology? *Quaternary Science Reviews*, *190*, 148–160.

Johansson, S. (2014). The thinking Neanderthals: What do we know about Neanderthal cognition? *Wiley Interdisciplinary Reviews. Cognitive Science*, *5*(6), 613–620.

Martínez, I., Rosa, M., Quam, R., Jarabo, P., Lorenzo, C., Bonmatí, A., Gómez-Olivencia, A., Gracia, A., & Arsuaga, J. L. (2013). Communicative capacities in Middle Pleistocene humans from the Sierra de Atapuerca in Spain. *Quaternary International*, *295*, 94–101.

Otte, M. (2019). Cognitive Capacities of the Neanderthals. In Y. Nishiaki & O. Jöris (Eds.), *Learning Among Neanderthals and Palaeolithic Modern Humans: Archaeological Evidence* (pp. 35–55). Singapore: Springer Nature.

Romagnoli, F., Rivals, F., & Benazzi, S. (2022). Updating Neanderthals: Taking stock of more than 160years of studies. In F. Romagnoli, F. Rivals, S. Benazzi (Eds.), *Updating Neanderthals: Understanding Behavioural Complexity in the Late Middle Palaeolithic* (pp. 1–15). Amsterdam: Academic Press.

Sykes, W. (2015). To see a world in a hafted tool: Birch pitch composite technology, cognition and memory in Neanderthals. In F. Coward, R. Hosfield, M. Pope, & F. Wenban-Smith, (Eds.), *Settlement, society and cognition in human evolution: Landscapes in the mind* (pp. 117–137). Cambridge: Cambridge University Press.

Slimak, L. (2019). For a cultural anthropology of the last Neanderthals. *Quaternary Science Reviews*, *217*, 330–339.

Stout, D., Apel, J., Commander, J., & Roberts, M. (2014). Late Acheulean technology and cognition at Boxgrove, UK. *Journal of Archaeological Science*, *41*, 576–590.

Artifacts, analogy, and metaphor: Inferring the cognitive foundations of metaphor from an archaeological and comparative perspective

Svetlana Kuleshova*[1,2], Elizabeth Qing Zhang[3], and Michael Pleyer[1]

*Corresponding Author: 40010189@parisnanterre.fr
[1]Center for Language Evolution Studies, Nicolaus Copernicus University in Toruń, Toruń, Poland
[2]ArScAn-Équipe AnTET (UMR 7041), CNRS, Université Paris Nanterre, Nanterre, France
[3]School of Linguistic Sciences and Arts, Jiangsu Normal University, Xuzhou, China

Metaphor has been shown to be a central process in human language and cognition (Lakoff & Johnson 1980). Moreover, it has also been assigned an important role in the evolution of language (Smith & Höfler 2015; Ellison & Reinöhl 2022). Uncovering the evolution of metaphor and the cognitive processes supporting it therefore presents an important part of explaining language evolution. Importantly, metaphorical cognition should not be seen as a unitary ability, but instead of as a multicomponent mosaic of underlying abilities that constitute it (Holyoak & Stamenković 2018). Such a 'decompositional' view has the advantage that the individual cognitive processes underlying metaphorical cognition and their evolutionary foundations can be investigated separately (Pleyer et al. 2023). This also has the advantage that the evolution of the cognitive foundations of metaphor can be traced with a deeper time depth than if treating it as a singular ability. Specifically, it allows us to investigate whether any of these abilities are evident to a degree in the behaviour of non-human animals, and whether they can be inferred from the archaeological record. Here, we focus on tool use as a source of evidence for the evolution of one central process supporting metaphor: analogy. We focus on analogy because "metaphors are predominantly relational comparisons, and are thus essentially analogies" (Gentner 1983). Specifically, we present two sources of evidence to investigate the evolution of analogy: archaeological and comparative data on tool use.

From the archaeological perspective, we propose to look for analogical abilities in the creation of stone tools, as it is widely accepted that analogy plays an important role in tool production and the invention process (Krumnack, Kühnberger, Schwering & Besold, 2020; Osiurak & Reynaud, 2020). Although it falls within the realm of cognitive archaeology, there are few examples of discussions of analogical capacities in prehistory (e.g., de Beaune 2004), and they concentrate on their evolution through different time periods. Here, we propose a methodology to look for analogical capacities in archaeological artifacts at a particular point in time. We suggest considering the productional diversity (i.e., different ways to achieve the same goal) of an archaeological collection. Differences in *chaînes opératoires* leading to the same productional goal may indicate the presence of problem-solving situations necessitating analogical capacities, as they presuppose the capacity to adapt known solutions to similar problems. Specifically, they do so based on analogical relations between a mental template representing a retrieval source on the one hand, and materials to be knapped or shaped, onto which inferences based on previous knowledge should be mapped, on the other. We develop this methodology using the examples of the Collection de la Pointe aux Oies, Wimeureux, France (Tuffreau, 1971) and the Collection de la Grande Vallée, Colombiers, France (Hérisson et al., 2016). The two collections differ in their modes of production: one consists of cores and flakes, and the other one of handaxes. The two examples will allow us to illustrate how our methodology can be implemented on different types of prehistoric tools.

From the perspective of comparative cognition, analogical abilities have also been found in tool use. For example, New Caledonian crows use two types of tools—hooked-twigs and stepped-cut tools—to achieve the same goal: looking for food in living and dead wood (Hunt, 1996). The manufacture of the hooked tools includes multiple steps with variations of material and ways of manufacturing (Hunt & Gray, 2003). Similarly, wild chimpanzees use leaves and moss as sponges to absorb water (Hobaiter et al., 2014), and their hands and folding leaves as "containers" to drink water (Sousa, Biro & Matsuzawa, 2009). They also crack nuts with a hammer-like tool on an anvil. The selection of the toolkit depends on multidimensional features, such as weight, material, distance to nut and the anvil (Sirianni, Mundry & Boesch, 2015). These data suggest that nonhuman animals can use different methods to achieve the same productional goal in an analogical fashion. Furthermore, there is also evidence for relational reasoning in nonhuman animals. Examples include honeybees, birds and nonhuman primates (Giurfa, 2021; Smirnova et al., 2021; Christie et al., 2016).

In sum then, we propose that investigating archaeological and comparative data on tool use and analogy can serve as a fruitful methodology to shed light on the evolution of metaphor and its underlying cognitive foundations.

Acknowledgements

SK thanks David Hérisson and Jean Airvaux, as well as the National Museum of Prehistory, for granting access to the collection of la Grande Vallée and giving their time to help her better understand this collection. SK also thanks Eric Boëda and the University of Paris Nanterre for granting access to the collection of la Pointe aux Oies. This research is part of the project No. 2021/43/P/HS2/02729 co-funded by the National Science Centre and the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 945339.

References

Christie, S., Gentner, D., Call, J., & Haun, D. B. M. (2016). Sensitivity to Relational Similarity and Object Similarity in Apes and Children. *Current Biology*, 26(4), 531–535.

De Beaune, S. (2004). The invention of technology: prehistory and cognition. *Current Anthropology*, *45*(2), 139-162.

Ellison, T. M., & Reinöhl, U. (2022). Compositionality, metaphor, and the evolution of language. *International Journal of Primatology*, 1-17.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2), 155-170.

Giurfa, M. (2021). Learning of sameness/difference relationships by honey bees: performance, strategies and ecological context. *Current Opinion in Behavioral Sciences*, 37, 1–6.

Hérisson, D. Airvaux, J., Lenoble, A., Richter, D., Claud, E., & Primault, J. (2016). Between the northern and southern regions of Western Europe: The Acheulean site of La Grande Vallée (Colombiers, Vienne, France). *Quaternary International*, *411* (B), 108-131.

Hobaiter, C., Poisot, T., Zuberbühler, K., Hoppitt, W., & Gruber, T. (2014). Social Network Analysis Shows Direct Evidence for Social Transmission of Tool Use in Wild Chimpanzees. *PLoS Biology*, 12(9), e1001960.

Holyoak, K. J., & Stamenković, D. (2018). Metaphor comprehension: A critical review of theories and evidence. *Psychological Bulletin*, 144(6), 641-671.

Hunt, G. R. (1996). Manufacture and use of hook-tools by New Caledonian crows. *Nature*, 379(6562), 249-251.

Hunt, G. R., & Gray, R. D. (2003). Diversification and cumulative evolution in New Caledonian crow tool manufacture. *Proceedings of the Royal Society B: Biological Sciences*, 270(1517), 867–874.

Krumnack, U., Kühnberger, K.-U., Schwering, A., & Besold, T. R. (2020). Analogies and analogical reasoning in invention. In E.G. Carayannis (Ed.), *Encyclopedia of Creativity, Invention, Innovation and Entrepreneurship: Second Edition* (pp. 56-63). New York: Springer.

Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By*. University of Chicago Press.

Osiurak, F., & Reynaud, E. (2020). The elephant in the room: What matters cognitively in cumulative technological culture. *Behavioral and Brain Sciences*, *43*.

Pleyer, M., Kuleshova, S., & Zhang, E. Q. (2023). Analogy and the Evolution of the Cognitive Foundations of Metaphor: A Comparative and Archaeological Perspective. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 45. Retrieved from https://escholarship.org/uc/item/2g59r2k8

Sirianni, G., Mundry, R., & Boesch, C. (2015). When to choose which tool: multidimensional and conditional selection of nut-cracking hammers in wild chimpanzees. *Animal Behaviour*, 100, 152–165.

Smith, A. D., & Höfler, S. (2015). The pivotal role of metaphor in the evolution of human language. In J. E. Díaz-Vera (Ed.), M*etaphor and Metonymy across Time and Cultures: Perspectives on the Sociohistorical Linguistics of Figurative Language*. De Gruyter.

Smirnova, A. A., Obozova, T. A., Zorina, Z. A., & Wasserman, E. A. (2021). How do crows and parrots come to spontaneously perceive relations-between-relations? *Current Opinion in Behavioral Sciences*, 37, 109–117.

Sousa, C., Biro, D., & Matsuzawa, T. (2009). Leaf-tool use for drinking water by wild chimpanzees (Pan troglodytes): Acquisition patterns and handedness. *Animal Cognition*, 12(1 SUPPL).

Tuffreau, A. (1971). Quelques Observations sur le Paléolithique de la Pointe-aux-Oies à Wimereux (Pas-de-Calais). *Bulletin de la Société préhistorique française. Études et travaux, 68*(2), 496-504.

Learnability effects in children: are more structured languages easier to learn?

Imme Lammertink[*1], Marianne De Heer Kloots[2], Mary Bazioni[3] and Limor Raviv[3,4]

[*]Corresponding Author: imme.lammertink@ru.nl
[1]CLS, Radboud University, Nijmegen, NL
[2]ILLC, University of Amsterdam, Amsterdam, NL
[3]LEADS group, Max Planck Institute for Psycholinguistics, Nijmegen, NL
[4]cSCAN, University of Glasgow, Glasgow, UK

Cross-linguistic differences in morphological complexity could have important consequences for language learning. A recent study by Raviv et al. (2021) showed that adults learn highly systematic artificial languages faster and more accurately than semi-, or non-structured artificial languages, suggesting that some languages may be acquired faster than others. However, these findings are limited in two ways. First, they are based only on adult learners, despite the fact that children are the most prototypical language learners in real-world situations and may differ in their learning biases from adults (e.g., Culbertson & Newport, 2015; Hudson Kam & Newport, 2005, 2009; Newport, 2020; Schuler, 2017; Tal & Arnon, 2022). Since children are often seen as the agents of language emergence (Senghas et al., 2004), their performance is thus a necessary test case for the hypothesis that languages with more regular, compositional, and transparent grammars are easier to learn, and that children may introduce more structural innovations during generalization. Second, it remains unclear how individual differences in learning-related cognitive capacities such as working memory and selective attention may impact these effects. For example, learners with better working memory might benefit less from the existence of more regularity, as they may be better able to remember all unique forms, without the need to rely on regularities. Addressing these limitations is important for refining theories on language evolution and the origin of linguistic diversity. Therefore, in this pre-registered study we extend previous findings to child learners, as well as test for the role of individual differences. The full pre-registration can be found at https://osf.io/w89ju.

Participants (105 adults and 105 children aged 8-11 years, all native Dutch speakers) first learn one of three child-friendly artificial languages, based on three artificial languages used in Raviv et al. (2021) (for details, see pre-registration). These child-friendly languages consist of 12 scene-label pairs, and vary in their level of compositional structure, i.e., the degree to which similar meanings were systematically expressed using similar strings: ranging from highly systematic languages (structure score 0.86), medium structured languages (structure score 0.67), to unstructured languages (structure score 0.36). After training, participants are tested on their knowledge of the language they learned and then complete an additional generalization test, where they are asked to produce labels for six new scenes not included in the training. Finally, we assess participants' working memory and selective attention using a Digit Span Backward test (Semel et al., 2010) and the Map Mission subtest (Manly et al., 2003; Robertson et al., 1994).

Data collection is still ongoing, but preliminary results from N=36 children and N=46 adults (Figure 1) show that adults outperform children during test and generalization – in line with previous artificial language learning experiments (Ferman & Karni, 2010; Perry et al., 2016; Raviv & Arnon, 2018). Both children and adults seem to show a learning and generalization advantage for more structured languages, supporting our prediction that the positive effect of systematicity is based on general principles of compressibility (Kirby, 2002; Zuidema, 2003) and should thus hold across one's lifespan. While working memory and selective attention may modulate this effect, we have not yet looked at individual differences given the relatively small size of the current dataset.

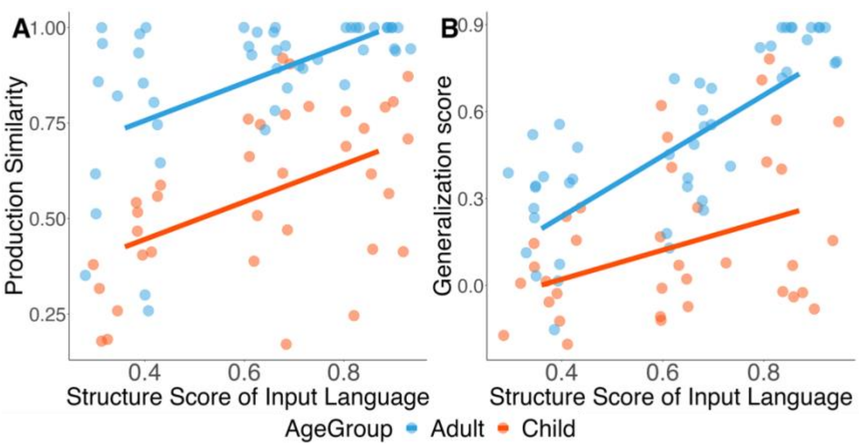


Figure 1 (A) production accuracy at test and (B) generalization scores for unseen items as a function of the language's structure score and age group. Each point represents the average of a single participant. The thick line represents the group average.

Acknowledgments
We thank Lotte Arendsen, Oxana Grosseck and Remi van Casteren for their help with recruitment and with testing adult and child participants.

References

Culbertson, J., & 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

Ferman, S., & 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

Hudson Kam, C. L., & 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., & 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

Kirby, S. (2002). Learning, bottlenecks and the evolution of recursive syntax. In Briscoe, T., editor, Linguistic Evolution through Language Acquisition: Formal and Computational Models, chapter 6, pages 173–204. Cambridge University Press.

Newport, E. L. (2020). Children and Adults as Language Learners: Rules, Variation, and Maturational Change. *Topics in Cognitive Science*, *12*(1), 153–169. https://doi.org/10.1111/tops.12416

Manly, T., Robertson, I., Anderson, V., & Nimmo-Smith, I. (2004). TEA-CH-NL Test of Everyday Attention for Children. [Measurement instrument] Amsterdam, The Netherlands: Pearson Clinical Assessment.

Perry, L. K., Axelsson, E. L., & Horst, J. S. (2016). Learning What to Remember: Vocabulary Knowledge and Children's Memory for Object Names and Features: Brief Report. *Infant and Child Development*, *25*(4), 247–258. https://doi.org/10.1002/icd.1933

Raviv, L., & 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., De Heer Kloots, M., & 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. https://doi.org/10.1016/j.cognition.2021.104620

Robertson, I., Ward, T., Ridgeway, V., & Nimmo-Smith, I. (1994). TEA-UK Test of Everyday Attention [Measurement instrument]. Pearson clinical assessment.

Schuler, K. (2017). *The Acquisition of Productive Rules in Child and Adult Language Learners* [Georgetown University]. https://www.proquest.com/docview/1958956127?parentSessionId=Je2hHEvolZsc3ryvwtzWnyQEs2OIjtByExg%2F9NVJBpY%3D

Semel, E., Wiig, E., & Secord, W. (2010). Clinical evaluation of language fundamentals: Dutch version (W. Kort, E. Compaan, M. Schittekatte, & P. Dekker, Trans.; Third Edition.) [Measurement instrument]. Amsterdam, The Netherlands: Pearson clinical assessment.

Senghas, A., Kita, S., & 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

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

Zuidema, W. (2003). How the poverty of the stimulus solves the poverty of the stimulus. Advances in Neural Information Processing Systems, 51–58.

Diversity and universals in culturally evolved sound systems across species: A complex adaptive system account

Li Lei[*1], Elizabeth Qing Zhang[2], and Tao Gong[3]

[*]Corresponding Author: leili@science.ru.nl
[1]Radboud University, Nijmegen, The Netherlands [2]Jiangsu Normal University, Xuzhou, China [3]Google, New York, USA

Diversity and universals are concurrent features in culturally evolved sound systems (Oudeyer, 2005). Diversity, as evidenced by cross-linguistic differences (Evans & Levinson, 2009), contrasts with universals, which here denote consistent statistical tendencies across languages (Bybee, 2010). For example, while vowel inventories vary significantly among languages (Maddieson, 1984), three prevalent vowels are consistently observed in the majority of languages (Schwartz, Boë, Vallée & Abry, 1997). Similarly, in cetacean vocal systems shaped by cultural transmission, although dialectal diversity exists in killer whale calls (Filatova et al., 2012) and sperm whale codas (Weilgart & Whitehead, 1997), certain calls in killer whales (Rehn, Filatova, Durban & Foote, 2011) and hierarchical structures in humpback whale songs (Payne & Payne, 1985) suggest universality, implying a potential convergent mechanism across species for sound system distribution and change in cultural evolution.

Research attempts to elucidate the origins of diversity and universals. The *Acoustic Adaptation Hypothesis* (Wiley & Richards, 1978), originally formulated in animal communications, extends to human spoken language (Maddieson & Coupé, 2015), highlighting diversity in adaptation to different ecological environments. Conversely, universals arise from cross-linguistic communicative tradeoffs (Coupé, Oh, Dediu & Pellegrino, 2019). Information per syllable and speech rate differ across languages, however, encodings for each language are largely balanced, efficient, and universal. Similar viewpoints have been proposed in other domains of language. The *Linguistic Niche Hypothesis* (Lupyan & Dale, 2016) emphasizes that morphological complexity reflects the structural complexity of speech communities, adapting to various social environments. Similarly, the *Efficient Communication Framework*

(Kemp, Xu & Regier, 2018) underscores the necessity for languages to balance informativeness (e.g., kinship terms of reference) and cognitive load (e.g., number of such terms) for high communicative efficiency.

This article aims to integrate these theories within a *complex adaptive systems* (CAS) account by comparing existing *agent-based models* (ABMs) of human vowel systems and cetacean vocal systems. These simulations explore evolutionary dynamics in CAS by simulating global changes through local interactions among artificial agents with cognitive capabilities (Holland, 2000).

De Boer's *Imitation Game* (2000) demonstrates how vowel universals can emerge through local interactions among agents without presupposing innate constraints. How adaptation leads to diversity was also investigated in some extensions of the model. De Boer and Vogt (1999) demonstrated that differences in social structures and demographics collectively shape vowel systems. Chirkova and Gong (2014) showed that when a new vowel entered into a vowel system, the original vowel distribution could dynamically adapt and readjust to maintain contrast with the original vowels and the newly-entered one, thus forming a novel vowel system.

It is generally accepted that cetacean vocalisation is learnt rather than transmitted genetically (Whitehead & Rendell, 2014). Research on cetacean vocalization suggests that dialectal diversity arises from female-centric social learning mechanisms and multilevel societies (Filatova & Miller, 2015; Cantor et al., 2015). The ABMs of humpback whale songs focused on factors that influence evolution patterns, such as migration and population contact (Mcloughlin et al., 2018). Although no model has explored which innate constraints shape the universal song form, research shows that similar learning parameters lead to different evolutionary patterns of the song for a universal goal (Zandberg, Lachlan, Lamoni & Garland, 2021).

These ABMs illustrate sound system evolution as a process of self-adaptive optimization. Diversity arises from differences in adaptation strategies across environments, while universals stem from common conditional constraints and specific goals (e.g., communicative for humans and biological for non-human animals, Steels, 2017). It can uncover generalizable principles underlying the emergence of diversity and universals in language evolution by comparing ABMs of sound systems. Such cross-species comparisons offer insights into the social-cognitive mechanisms driving language evolution and the perpetuation of endless change (Steels, 2017). Moreover, similar issues of diversity and universals exist in other aspects of human languages (e.g., color terms, Gong et al., 2019), suggesting that these principles may extend beyond sound systems.

Acknowledgements

We would like to thank three anonymous reviewers for their insightful suggestions and comments.

References

Bybee, J. (2010). *Language, usage and cognition*. Cambridge University Press.

Cantor, M., Shoemaker, L. G., Cabral, R. B., Flores, C. O., Varga, M., & Whitehead, H. (2015). Multilevel animal societies can emerge from cultural transmission. *Nature Communications, 6*(1), 8091.

Chirkova, K., & Gong, T. (2014). Simulating vowel chain shift in Xumi. *Lingua*, 152, 65-80.

Coupé, C., Oh, Y. M., Dediu, D., & Pellegrino, F. (2019). Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science Advances, 5*(9), eaaw2594.

De Boer, B. (2000). Self-organization in vowel systems. *Journal of Phonetics, 28*(4), 441-465.

De Boer, B., & Vogt, P. (1999). Emergence of speech sounds in changing populations. In D. Floreano, J. D. Nicoud, & F. Mondada (Eds.), *Advances in Artificial Life*. Springer.

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32(5), 429-448.

Filatova, O. A., Deecke, V. B., Ford, J. K., Matkin, C. O., Barrett-Lennard, L. G., Guzeev, M. A., Burdin A. M., & Hoyt, E. (2012). Call diversity in the North Pacific killer whale populations: Implications for dialect evolution and population history. *Animal Behaviour, 83*(3), 595-603.

Filatova, O. A., & Miller, P. J. (2015). An agent-based model of dialect evolution in killer whales. *Journal of Theoretical Biology, 373*, 82-91.

Holland, J. H. (2000). *Emergence: From chaos to order*. Oxford University Press: Oxford.

Gong, T., Gao, H., Wang, Z., & Shuai, L. (2019). Perceptual constraints on colors induce the universality of linguistic color categorization. *Scientific Reports*, 9, 7719.

Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics, 4*, 109-128.

Lupyan, G., & Dale, R. (2016). Why are there different languages? The role of adaptation in linguistic diversity. *Trends in Cognitive Sciences, 20*(9), 649-660.

Maddieson, I. (1984). The effects on F0 of a voicing distinction in sonorants and their implications for a theory of tonogenesis. *Journal of Phonetics, 12*(1), 9-15.

Maddieson, I., & Coupé, C. (2015). Human spoken language diversity and the acoustic adaptation hypothesis. *The Journal of the Acoustical Society of America, 138*(3), 1838-1838.

Mcloughlin, M., Lamoni, L., Garland, E. C., Ingram, S., Kirke, A., Noad, M. J., Rendell, L., & Miranda, E. (2018). Using agent-based models to understand the role of individuals in the song evolution of humpback whales (*Megaptera novaeangliae*). *Music & Science, 1*, 2059204318757021.

Oudeyer, P. Y. (2005). From holistic to discrete speech sounds: The blind snowflake-maker hypothesis. In M. Tallerman (Ed.), *Language origins: Perspectives on evolution*, 68-99.

Payne, K., & Payne, R. (1985). Large scale changes over 19 years in songs of humpback whales in Bermuda. *Zeitschrift für Tierpsychologie, 68*(2), 89-114.

Rehn, N., Filatova, O. A., Durban, J. W., & Foote, A. D. (2011). Cross-cultural and cross-ecotype production of a killer whale 'excitement' call suggests universality. *Naturwissenschaften, 98*, 1-6.

Schwartz, J. L., Boë, L. J., Vallée, N., & Abry, C. (1997). Major trends in vowel system inventories. *Journal of Phonetics, 25*(3), 233-253.

Steels, L. (2017). Human language is a culturally evolving system. *Psychonomic Bulletin & Review, 24*, 190-193.

Weilgart, L., & Whitehead, H. (1997). Group-specific dialects and geographical variation in coda repertoire in South Pacific sperm whales. *Behavioral Ecology and Sociobiology, 40*, 277-285.

Whitehead, H., & Rendell, L. (2014). *The cultural lives of whales and dolphins*. University of Chicago Press.

Wiley, R. H., & Richards, D. G. (1978). Physical constraints on acoustic communication in the atmosphere: implications for the evolution of animal vocalizations. *Behavioral Ecology and Sociobiology, 3*, 69-94.

Zandberg, L., Lachlan, R. F., Lamoni, L., & Garland, E. C. (2021). Global cultural evolutionary model of humpback whale song. *Philosophical Transactions of the Royal Society B, 376*(1836), 20200242.

Obsolescence: a missing piece in the puzzle of polysemy and time

Vivian G. Li*

*liguo.vivian@gmail.com
*Department of Linguistics, Yale University, New Haven, USA

Previous studies reported a positive correlation between a word's age and the number of senses it has. The underlying mechanism has been attributed to the idea that semantic changes inevitably occur over time (i.e., words do not stop changing in meaning even when it already has several senses). However, evidence from those studies falls short of fully supporting what the purported mechanism entails, as those studies focused exclusively on non-obsolete senses in non-obsolete words. A comprehensive test of the predictions of the diachronic mechanism requires evidence that the reported positive correlation extends to obsolete words, and that the correlation holds when obsolete senses are also counted. This study provides that missing piece for the puzzle. Examining over 36,000 English verbs, we show that longer lifespan correlates with a greater number of changes/senses in obsolete as well as non-obsolete words, though the effect is stronger in non-obsolete words.

1. Introduction

As a language evolves, the meanings of its words change, which often involves gaining or losing senses. While there have been extensive discussions on the nature and representation of polysemy (Robins, 1967; Geeraerts, 1993; Norvig & Lakoff, 1987), relatively less is known about its diachronic development, e.g. do words inevitably undergo more changes over time?

This question has been addressed in some studies, but results remain, in our opinion, inconclusive. Lee (1990) hypothesized that words that have entered the language early would be more polysemous than words that have entered the language more recently. The logic being that *older words have more time to undergo changes than younger words*. Analyzing three sets of about 200 randomly sampled English nouns and adjectives, Lee showed that year of entry indeed negatively correlated with the number of (non-obsolete) senses a word has. Flieller and Tournois (1994) similarly reported a significant correlation between year of entry and the status of being polysemous (as opposed to monosemous) in a sample of 998 French words consisting of nouns, verbs, adjectives and adverbs. Berdicevskis (2020) extended Lee (1990)'s analysis to all non-obsolete nouns, adjectives and verbs documented in the Oxford English Dictionary, and concurred with Lee (1990)'s conclusions.

We consider the evidence from Lee (1990) and subsequent studies inconclusive because it falls short of fully supporting what the purported mechanism entails. The underlying diachronic mechanism (i.e., time begets changes) amounts to claiming that "all words would undergo more *changes* (or, more *sense-gaining events*) over time," which is a prediction about obsolete as well as non-obsolete words, and is a prediction about the total number of changes (or the number of sense-gaining events) rather than the net amount of senses that survived till the present day. Lee (1990) and subsequent studies focused exclusively on non-obsolete senses in non-obsolete words, thus leaving their findings susceptible of survivorship bias.

A comprehensive test of the predictions of the diachronic mechanism therefore requires evidence that the reported positive correlation between a word's age and the number of senses/changes extends to obsolete words, and that the correlation holds when obsolete senses are counted. If, for instance, obsolete words are generally monosemous irrespective of the length of their lifespans, then the positive correlation would hold only for non-obsolete words; consequently, the correlation should not be taken as a diachronic generalization about polysemy. Similarly, if words that entered the language recently appear less polysemous (i.e., possess fewer non-obsolete senses) than words that entered early because they have disproportionally many obsolete senses, then the correlation is likely a corollary of older words having a higher proportion of non-obsolete senses than younger words do, rather than a diachronic regularity of polysemy.

In what follows, we provide the missing piece in the puzzle of polysemy and time, by extending Lee (1990) and Berdicevskis (2020)'s analyses to obsolete words and obsolete senses.

2. Method

Our primary source of data is verbs from the online version of the Oxford English Dictionary (OED). We focus on verbs because verbs tend to show more changes than nouns or adjectives (Dubossarsky, Weinshall, & Grossman, 2016).

OED organizes word senses into hierarchical groups, such that a word may have several major sense categories, and under each major sense category, there are sub-senses. Most of the subsenses are dated. We counted the dated senses as a proxy for the number of senses, on the assumption that if the senses could be dated (potentially differently), that in itself is evidence that those senses are distinct (see e.g. Hamilton, Leskovec, and Jurafsky (2016) on the difficulty in counting senses). The number of changes a word has gone through is counted as the sum of 1) the number of non-obsolete senses and 2) twice the number of obsolete senses, to take both gain and loss of senses into account.

Lifespan of a word is calculated as the span between the earliest and the latest year of use among all senses. If any of a word's senses are still in use, the latest year would be 2023. If the earliest year of use is "Old English" or "late Old English", we use 800 and 1025 as a proxy, as they are roughly the midpoint of those respective periods (Baugh & Cable, 1993; Sweet, 1990). If the earliest year of use is annotated with "*a*(nte)", "*c*(irca)", or ".." (e.g. *17..*), only the numeric value would be used for calculating the lifespan (e.g. *a1700, c1700, 17..* would all be treated as *1700*). If a word's lifespan thus calculated comes out as 1 (meaning the earliest and latest year of use are the same), and the earliest year of use contained annotations (e.g., "Old English"), the word would be excluded from analysis, since the estimated lifespan would likely deviate substantially from the actual value.

For nonobsolete words, we also extracted their frequency band value (this information is unavailable for obsolete words),[1] which reflects the words' overall frequency in written English from 1970 to the present. The frequency bands denote nine levels of frequencies (on a logarithmic scale),[2] with 8 being the most frequent ($>$ 1000 per million words), and 0 being the least frequent.

Where appropriate, we ran parallel analyses on a set of non-obsolete verbs, nouns, and adjectives extracted from the same source in 2019 (Berdicevskis, 2020).

3. Results

Both obsolete and nonobsolete verbs show an inverse relationship between the number of words with n changes/senses and *n* (Fig. 1). More than half of the words (14416 out of 26480 nonobsolete verbs; 7864 out of 9808 obsolete verbs) are monosemous. Correspondingly, those words have only minimal number of changes: one change (sense gain) for non-obsolete words, two changes (gain and loss) for obsolete words. The most polysemous non-obsolete verb (*set*) has undergone 335 changes, with a total of 262 senses, whereas the most polysemous obsolete verb (*yknow*) has undergone 38 changes, with 19 senses in total.

To examine the effect of lifespan and potential differences between obsolete and nonobsolete words (Fig. 2), we fit a Poisson regression model on the entire dataset, with the *number of changes* as the dependent variable, and *lifespan*, word's *state* (obsolete vs. nonobsolete), and their *interaction* as independent variables. Results (Table 1) revealed that a word is more likely to have a high number of changes when it has a longer lifespan (for nonobsolete words: $\beta_{lifespan} = 0.0037$; for obsolete words: $\beta_{lifespan} = 0.0013$) , although the effect is weaker in obsolete words than in nonobsolete words ($\beta_{lifeSpan \times state} = -0.0024, p < 0.0001$).

[1]OED distinguishes obsolete and non-obsolete words. A word is obsolete if all its senses are obsolete, "this usually means that no evidence for the term can be found in modern English" (Oxford University Press, n.d.b).

[2]Although the description on the official site (https://www.oed.com/information/understanding-entries/frequency/) states that there are eight levels of frequency band, some non-obsolete words are marked with 0 ("Frequency: 0 out of 8" in the html code). We therefore distinguish nine levels of frequencies.

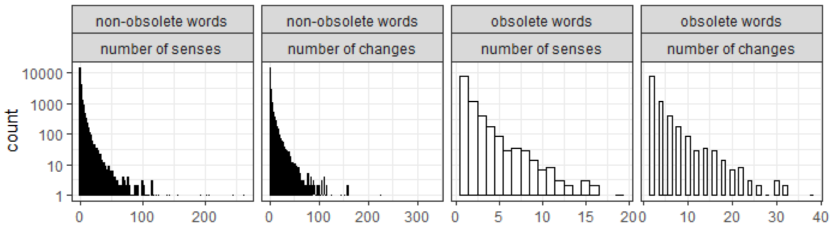


Figure 1. Histograms for the distribution of non-obsolete and obsolete words by the number of changes/senses they have. Y axis is on logarithmic scale.

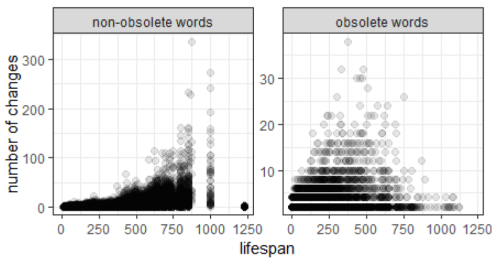


Figure 2. Distribution of lifespan by the number of changes, in non-obsolete and obsolete words. Each grey circle represents a data point (i.e. a word). Dark regions indicate overlapping data points.

Table 1. Coefficients from the regression model on the number of changes.

Predictor	Coefficient	SE	z value	p value
(Intercept)	-0.1240	0.0066	-18.65	< 0.0001
lifespan	0.0037	0.00001	338.66	< 0.0001
state=obsolete	0.9615	0.0102	94.42	< 0.0001
lifespan $\times$ state=obsolete	-0.0024	0.00003	-78.42	< 0.0001

A Poisson regression model using *lifespan*, *word state* (obsolete vs. nonobsolete), and their *interaction* to predict the total number of senses showed a similar pattern. Results echo the findings on the number of changes: a word is more likely to have more senses when it has a longer lifespan (for nonobsolete words: $\beta_{lifespan} = 0.0034$, for obsolete words: $\beta_{lifespan} = 0.0013$, although the effect is weaker in obsolete words than in nonobsolete words ($\beta_{lifespan \times state} = -0.0021, p < 0.0001$).

To check if the addition of frequency alters the results, we fit a Poisson regression model on non-obsolete verbs alone, with the *number of changes* as the dependent variable, and *lifespan*, *frequency-band*, and their *interaction* as independent variables. Since frequency band values reflect words' frequency differences on a logarithmic scale, we treat them as a continuous variable (c.f. Berdicevskis, 2020). Lifespan ($\beta = 0.0012, p < 0.0001$) remains a significant positive predictor for the number of changes, so are frequency band ($\beta = 0.2638, p < 0.0001$) and their interaction ($\beta_{lifespan \times freqBand} = 0.0003, p < 0.0001$). These results indicate that among non-obsolete verbs, a word is more likely to have a high number of changes if it is of a long lifespan or high frequency; the effect of lifespan is enhanced in high frequency words in comparison to low frequency words (or alternatively, the effect of frequency is more prominent among words with longer lifespan than among words of shorter lifespan). A Poisson regression model for the total number of senses in nonobsolete verbs showed the same pattern.

To gauge if our findings would generalize to other parts of speech, we fit Poisson regression models to the dataset used in Berdicevskis (2020). We regressed the number of changes/senses (total_nmeanings) against year of entry (with the earliest year, 950, re-set as 0, to make the intercept more interpretable, as in Berdicevskis, 2020), frequency band, part of speech, and all two-way and three-way interactions. The effects of year and frequency are consistent with our findings based on verbs alone (see Fig. 3 for an illustration of $year(lifespan) \times frequency$ interaction).

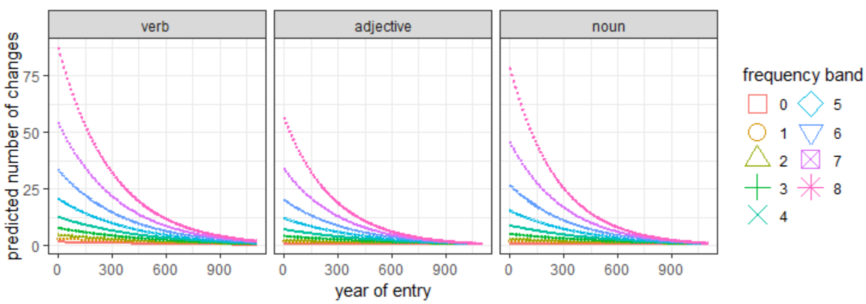


Figure 3. Predicted number of changes from the model based on Berdicevskis (2020)'s data.

4. Discussion

In this study we explored the relation between time and meaning change. Our analysis improves over previous ones in that ours incorporated obsolete words and obsolete senses, which are necessary components for assessing diachronic generalizations but have long been overlooked. We showed that longer lifespan is in general positively correlated with more changes in meaning (and the total number of senses), for non-obsolete and obsolete words alike, although the effect of lifespan is stronger in non-obsolete words than in obsolete words. Our finding complements Berdicevskis (2020), Lee (1990) and others and strengthens the claim that *words undergo more changes (or become more polysemous) over time*.

Our analyses also offer valuable insights to word obsolescence and the distribution of polysemy. Firstly, it has been claimed that “polysemy is pervasive” (Falkum & Vicente, 2015) and that “virtually every word is polysemous to some extent” (Vicente & Falkum, 2017), but such claims have not been empirically examined. Our results suggest that polysemy may not be as pervasive as expected, or rather, a large portion of the lexicon may be reasonably considered monosemous (for instance, as having one core meaning liable to contextual variation rather than having several more disparate senses). Our analyses are based on the OED ; future research could explore the empirical distribution of monosemy vs. polysemy with alternative approaches to track word/sense obsolescence and quantify polysemy. Secondly, in addition to the difference in the magnitude of lifespan effects, we observed a sharp contrast between obsolete and non-obsolete words in terms of the maximum number of senses (19 vs. 262). It is not clear if 19 represents an upper bound for the number of senses obsolete words could have. Future studies could delve into the nature of word obsolescence, and delineate how obsolete and non-obsolete words differ in other respects. Thirdly, we identified a Zipfian-like (Zipf, 1932) relation between the number of words with n senses and n. Subsequent investigations could survey the distribution of word senses in other languages, and look into the origins of this pattern.

In our analyses of obsolete words, we did not have their frequency information, and consequently excluded frequency as a predictor. The finding of significant lifespan effects in models without frequency indicates that lifespan may be a general effect in words of all frequencies. For comparison, we included frequency-band in models for non-obsolete words alone. Although frequency has been shown to correlate with semantic change and polysemy (Hamilton et al., 2016; Zipf, 1945), recent studies argued that frequency is not a causal factor for semantic change (Dubossarsky et al., 2016; Dubossarsky, Weinshall, & Grossman, 2017; Keidar, Opedal, Jin, & Sachan, 2022) and that change in meaning precedes changes in frequency (Feltgen, Fagard, & Nadal, 2017). Corroborating Berdicevskis (2020), frequency-band in our models correlated significantly with the number of changes/senses, and its interaction with lifespan was significant. We hypothesize that frequency, if it is not a causal factor, may have functioned as a proxy for some covert factor, e.g. prototypicality of senses (Dubossarsky et al., 2016). Note that in those models, the effect of lifespan was not subsumed by frequency, which strengthens the validity of lifespan as an independent predictor for the number of changes/senses. Irrespective of how frequency effects are interpreted, the significant interaction between lifespan and frequency implies, at least, that subclasses of words may exhibit diverse diachronic trajectories. Future studies may explore further how subclasses of a part of speech could vary systematically in their diachronic behaviour.

References

Baugh, A., & Cable, T. (1993). *A history of the English language.* Routledge.

Berdicevskis, A. (2020). Older English words are more polysemous. In A. Ravignani, C. Barbieri, M. Flaherty, Y. Jadoul, E. Lattenkamp, H. Little, M. Martins, K. Mudd, & T. Verhoef (Eds.), *The Evolution of Language: Proceedings of the 13th International Conference (Evolang13)* (p. 14-21).

Dubossarsky, H., Weinshall, D., & Grossman, E. (2016). Verbs change more than nouns: a bottom-up computational approach to semantic change. *Lingue e linguaggio*, *15*(1), 7–28.

Dubossarsky, H., Weinshall, D., & Grossman, E. (2017). Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 1136–1145).

Falkum, I. L., & Vicente, A. (2015). Polysemy: Current perspectives and approaches. *Lingua*, *157*, 1-16.

Feltgen, Q., Fagard, B., & Nadal, J.-P. (2017). Frequency patterns of semantic change: corpus-based evidence of a near-critical dynamics in language change. *Royal Society open science*, *4*(11), 170830.

Flieller, A., & Tournois, J. (1994). Imagery value, subjective and objective frequency, date of entry into the language, and degree of polysemy in a sample of 998 French words. *International Journal of Psychology*, *29*(4), 471–509.

Geeraerts, D. (1993). Vagueness's puzzles, polysemy's vagaries. *Cognitive Linguistics*, *4*, 223-272.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistic* (p. 1489-1501). Stroudsburg, PA: Assoc. Comput. Linguist.

Keidar, D., Opedal, A., Jin, Z., & Sachan, M. (2022). Slangvolution: A causal analysis of semantic change and frequency dynamics in slang. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (p. 1422–1442). Dublin, Ireland: Association for Computational Linguistics.

Lee, C. J. (1990). Some hypotheses concerning the evolution of polysemous words. *Journal of Psycholinguistic Research*, *19*, 211–219.

Norvig, P., & Lakoff, G. (1987). Taking: A study in lexical network theory. In J. Aske, N. Beery, L. A. Michaelis, & H. Filip (Eds.), *Annual meeting of the Berkeley Linguistics Society* (Vol. 13, pp. 195–206). Berkeley: Berkeley Linguistics Society.

Oxford University Press. (n.d.a). *Oxford English Dictionary.* `https://www.oed.com/`. (Accessed: 2023-09-30)

Oxford University Press. (n.d.b). *OED terminology.* `https:`

//www.oed.com/information/understanding-entries/oed-terminology. (Accessed: 2024-01-08)

Robins, R. H. (1967). *A short history of linguistics.* Bloomington: Indianna University Press.

Sweet, H. (1990). *A new English grammar: Logical and historical. part 1: Introduction, phonology and accidence* (Vol. 1). Oxford At The Clarendon Press.

Vicente, A., & Falkum, I. L. (2017). Polysemy. In *Oxford research encyclopedia of linguistics.* Retrieved 11 Oct. 2023, from https://oxfordre.com/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-325.

Zipf, G. K. (1932). *Selected studies of the principle of relative frequency in language.* Harvard university press.

Zipf, G. K. (1945). The meaning-frequency relationship of words. *The Journal of general psychology*, *33*(2), 251–256.

Why are Some Words More Frequent than Others? New Insights from Network Science

Qiawen Liu*[1], Simon De Deyne[2], and Gary Lupyan[1]

*Corresponding Author: qliu295@wisc.edu
[1]Department of Psychology, University of Wisconsin-Madison, Madison, United States
[2]School of Psychological Sciences, University of Melbourne, Melbourne, Australia

Why are some words more frequent than others? Surprisingly, the obvious answers to this seemingly simple question, e.g., that frequent words reflect greater communicative needs, are either wrong or incomplete. We show that a word's frequency is strongly associated with its position in a semantic association network. More centrally located words are more frequent. But is a word's centrality in a network merely a reflection of *inherent* centrality of the word's meaning? Through cross-linguistic comparisons, we found that differences in the frequency of translation-equivalents are predicted by differences in the word's network structures in the different languages. Specifically, frequency was linked to how many connections a word had and to its capacity to bridge words that are typically not linked. This hints that a word's frequency (and with it, its meaning) may change as a function of the word's association with other words.

1. Introduction

Word frequencies are often used as a key predictor in studies of word recognition (Brysbaert et al., 2016; Ferrand et al., 2010; Keuleers et al., 2012), comprehension (Just & Carpenter, 1980; Halgren & Smith, 1987), production (Oldfield & Wingfield, 1965; Jescheniak & Levelt, 1994; Alario et al., 2004), recall (Arndt & Reder, 2002; Clark, 1992; Gregg, 1976; Meier et al., 2013; Yonelinas, 2002), and learning (Braginsky et al., 2019). While much is known about what word frequency predicts, much less is known about what predicts word frequencies (e.g., Calude & Pagel, 2014, 2011; Liu et al., 2023). Why are some words more frequent than others? Why do some words become more frequent while others become less frequent over time? How similar are frequencies of translation equivalents across languages and what does it mean if a word denoting a certain meaning is more frequent in one language than in another?

The question of why some words are more frequent than others suggests some obvious answers. One is that more frequent words denote meanings that are more important for people's goals and needs. 'Water' is more frequent than 'lamp','matrix', or 'abracadabra' because it is more important. This explanation only goes so far, however because important meanings are fragmented into more basic terms. Presumably 'mammals' are more important than 'dogs' or 'cats' yet the frequency of ‘mammal’ is a small fraction of either of those more basic terms (in fact ‘dog’ is about as frequent as ‘animal’). Another possibility is that more prototypical referents have higher frequencies (e.g., ‘robin’ compared to ‘penguin’) because they more closely correspond with what speakers have in mind (Rosch et al., 1976). However, what counts as a prototype can vary significantly based on context, making prototypicality an inconsistent predictor of word frequency. For example, prototypicality as a bird may explain why ‘robin’ is more frequent than ‘penguin’, but not why ‘chicken’ is more frequent than ‘robin’. It is also possible that word frequencies might mirror the prevalence of certain objects in our surroundings. But discrepancies arise here too. ‘Red’ is the most frequent chromatic term even though red objects are not more common. And explanations invoking ecological frequencies cannot explain the frequencies of abstract words that refer to intangibles. Explanations that stress communicative needs: ‘frequent words denote things we most want to talk about’ also run into problems. First, they simply push the question of word meanings to communicative need. ‘Girl’ is more frequent than ‘boy’, but do we really have a greater need to communicate about girls than boys? All these explanations also struggle with explaining why words with similar meanings are more frequent in some languages than others.

In a recent study, Liu et al. (2023) predicted word frequencies from properties of the words’ semantic networks. To rule out idiosyncratic explanations such as importance and ecological frequency, they examined pairs of antonyms which would seem to have equal communicative value but often differ in word frequency (as the example of girl/boy above). After factoring out effects like morphological complexity and polysemy, the analysis revealed two network properties that predicted word frequency especially well: the number of connections the word and its associated words have, and the word’s ability to bridge otherwise sparsely linked words. As further revealed by a longitudinal analysis, these network properties didn’t seem to just correlate with current word frequencies. Instead, they also predicted the way the word’s frequency changed in the subsequent decades, suggesting a causal role of network properties in explaining changes to word frequencies over time.

One alternative explanation is that the more frequent words in each antonym pair correspond to a default or unmarked state (Clark, 1992). For instance, ‘good’ in the pair good-bad is unmarked such that asking ‘how good was it?’ does not imply goodness, while asking ‘how bad was it?’ implies badness. Markedness-focused explanations make a simple prediction: if one end of a semantic dimension denoted by an antonym pair is inherently more central to communication and/or thinking (which leads to greater frequency of the associated word), then translation equivalents of these antonym pairs should show consistent frequency differences across languages. In the next section, we examined if translation equivalents of antonym pairs in Chinese and English display analogous frequency patterns. We then sought to replicate our earlier findings (2023) concerning the influence of network properties on word frequency using Chinese word-association data.

2. Effects of network centrality on Chinese word frequencies

2.1. *Materials*

We used the English and Chinese semantic association networks from the Small World of Words (SWOW) project. In this project, crowdsourced word association responses were gathered in various languages (De Deyne et al., 2019). Participants were shown target words and asked to list the first three words that came to mind. These associations then acted as cues for subsequent participants, generating further associations. This iterative method yielded a weighted network with directed edges. The edge direction signifies forward or backward associations, while the weight represents the likelihood of each association based on the response or cue. To focus on the most robust associations, we used only the first response and excluded responses provided by only a single respondent. The Chinese SWOW network had 21434 words and 78057 directional associative links.

In Liu et al. (2023), 774 antonym pairs of English words were extracted from WordNet (Fellbaum, 1998). Among them, 661 pairs were considered having appropriate Chinese translations, and were translated by a professional translator, resulting in 761 Chinese translation equivalents as some words had multiple valid translations. Word frequencies for both English and Chinese pairs were sourced from the Exquisite Corpus using the ‘wordfreq‘ Python package (Speer, 2022).

2.2. *Variables*

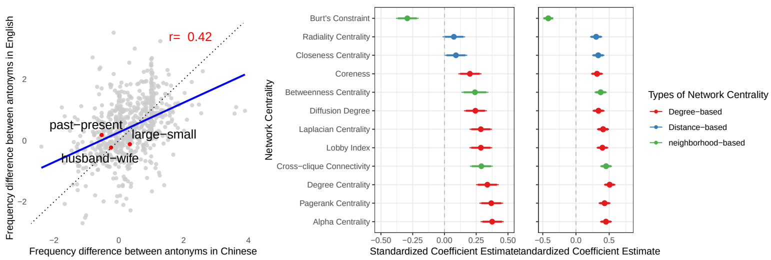
Using a linear regression model, we predicted the difference in Zipf frequency (calculated as the base-10 logarithm of occurrences per billion words) from different types of network centrality measures. These centralities are grouped into *degree-based* centralities which emphasize the number of connections a word and its neighbors have, *neighborhood-based* centralities which measure how well a word bridges between less-connected words, and *distance-based* centralities which consider words with short paths to others in the network as more central. We also included three covariates: the difference in morpheme count (more complex derived words may be less frequent), the differences in number of word senses (operationalized as Chinese Wordnet synsets) (Wang & Bond, 2013), and how often the word was mentioned as a cue, i.e., its frequency in SWOW. This allows us to discern the impact of the word's network above and beyond the frequency effect that some centrality measures may inevitably capture.

2.3. *Analysis & Result*

First, we examined the word frequency patterns between English and Chinese. As shown in Figure 2a, there's a moderate correlation ($r = .42$) between English-

Chinese differences in antonym pairs. This indicates that although there's a relationship in frequency patterns between the two languages, the prominence of a word in one language doesn't necessarily denote its inherent significance in meaning. For example, while 'small' is more frequent than 'large' in English, the Chinese counterpart xiǎo (small) is less frequent than dà (large); Do such frequency variations align with the word's network properties?

Because network centralities are highly inter-correlated, we regressed the difference in word frequency on each network centrality individually. We also controlled for differences in morpheme count, sense count, and SWOW frequency. In cases where an English antonym pair matched multiple Chinese antonym pairs, we averaged the measures for various translated pairs. Some words were absent from the Chinese WordNet or Chinese SWOW network, leaving us with 381 pairs for the analysis. We log-transformed notably skewed predictors. Figure 1b shows that, after controlling for other variables, degree-based, neighborhood-based, and distance-based network centralities all significantly predict frequencies (all significant at $\alpha = .01$, except closeness which significant at $\alpha = .05$). Radiality is a marginally significant predictor ($p = .1$). The predictions were all in the expected direction, with the only negative coefficient for Burt's constraint indicating that words with fewer redundant neighbors have higher frequencies, replicating Liu et al. (2023)'s results with the English SWOW network using the Chinese SWOW network. Words that are more associated with others, closer to others, and bridge otherwise less connected words are more frequent.

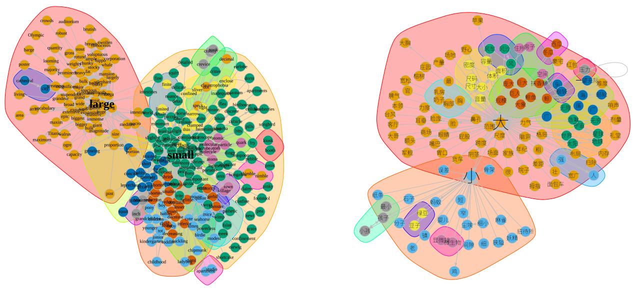


(a) (b) (c)

Figure 1.: (a) Which end of an antonym pair is more frequent is only moderately correlated between English and Chinese, e.g., 'past-present' favors 'present' in English and 'past' in Chinese. 'large-small' favors 'small' in English 'large' in Chinese. (b) Network centrality measures significantly predict Chinese word frequencies. (c) The differences between Chinese and English word frequencies for matched word pairs are predicted by differences in centrality measures.

3. Do cross-linguistic differences in network centralities predict cross-linguistic differences in word frequencies?

Are differences in word frequencies between English and Chinese associated with differences in the word's respective semantic networks? We predicted cross-linguistic differences in word frequencies between English and Chinese antonyms from the cross-linguistic differences in network centralities between English and Chinese antonyms, controlling for the same covariates as above. As shown in Figure 1 c, all cross-linguistic differences in centralities remain significant predictors of cross-linguistic differences in word frequency ($p < .01$).



(a) (b)

Figure 2.: Subnetworks of a translation-equivalent word pair. In English (a), *small* is more frequent than *large* and is also connected to more words & diverse neighborhoods (colored clusters). (b) In Chinese *dà, large* is more frequent and central than *xiǎo, small*.

4. Discussion

We analyzed the cross-linguistic frequency differences between English and Chinese antonyms, replicating the previous finding from English, in Chinese. The results reveal that words with greater degree-based, neighborhood-based, and distance-based centrality are, on average, more frequent. Specifically, words with more neighbors, especially more influential neighbors as determined by measures like PageRank or alpha centrality (which assign weights based on the importance of neighbors), tend to be more frequent (Fig 1 b). Words bridging less connected areas (Burt's constraint, Cross-clique connectivity, Betweenness centrality) and those with shorter paths to other words (closeness centrality) also tend to be more frequent. Moreover, the frequency differences across these languages can be attributed to differences in the words' network centrality. Overall, we show that variations in word frequency can be linked to the structural properties of the semantic network rather than solely to the inherent conceptual prominence of their denoted meanings. Differences in association network dynamics may underpin language evolution and influence patterns of word usage across languages.

How do cross-linguistic differences in network centralities inform our understanding of cultural variations in word meanings? Returning to the previously used example of cross-linguistic differences between large and small, we show that the more frequent word in both cases, despite having opposite meanings, is more centrally located and is a better 'bridge' to other meanings (Fig 2). One hypothesis linking centrality and frequency is that a more centrally located word has a higher base-level of activation during speech comprehension and production due to receiving more input from its neighbors, leading to a greater likelihood of a user producing it–a rich get richer type phenomenon.

Future research may further elucidate how fluctuations in network connectivity and cognitive accessibility influence the dynamics of competing synonyms (Karjus et al., 2020). In addition, words that connect otherwise less connected neighborhoods indicate they may have higher contextual diversity and larger semantic extensions compared to words that are surrounded by more redundant interconnected neighbors. Again, as shown in Figure 2, 'small' in English is more frequent and associates with a more diverse set of neighbors than 'large'; while this pattern is reversed in Chinese. For instance, the English 'small' is incorporated into phrases like 'small talk', helping to increase its opportunities for use compared to 'large' which lacks similar sense-extension. Conversely, in Chinese 'dà' (large) can refer to generality and lack of precision, as in 'dà gài' (probably or approximately), 'dàjú' (overall situation), 'dàyì' (careless), while 'xiǎo' (small) does not have analogous extension to higher certainty or precision.

Finally, what causes a word to occupy a more or less central location in a semantic network? Studies suggest that the structure of semantic networks is shaped by both external interactions with the environment (Hills et al., 2009b; Laurino et al., 2023) and internal linguistic structures (Steyvers & Tenenbaum, 2005). For example, the 'preferential acquisition hypothesis' (Hills et al., 2009a) suggests that children learn new words based on their prevalence and connections within the surrounding environment, independent of their existing vocabulary. On the other hand, the 'preferential attachment' theory (Steyvers & Tenenbaum, 2005) suggests that new words are more likely to be integrated into a child's vocabulary if they connect to already well-connected words, reinforcing the significance of these central nodes. Given that individual-level cognitive selection can predict global language change (Li et al., 2024), a word's centrality might stem from its relevance in real-world contexts and its connectivity within the language structure.

Many questions remain. What is the causal direction between network centrality and frequency we observe here? Although Liu et al., 2023 found that changes in network centrality predicted subsequent changes in the words' frequencies, the changes in frequencies of the words in the sample were quite small for the tested time period. More insight can be gained from analysis of words that have undergone rapid changes in frequency. These are sometimes accompanied by shifts of meaning: The frequency of 'broadcast' hugely increased when its meaning shifted from sowing seeds by scattering, to radio and TV transmissions. It is difficult to tell how its semantic network changed at that time, but it is possible to study the semantic networks of words currently undergoing rapid changes in frequency such as those studied by Grieve, Nini, and Guo (2017). It is also possible to experimentally manipulate a word's location in a participant's semantic network to see if it causes changes in likelihood of production.

References

Alario, F. X., Ferrand, L., Laganaro, M., New, B., Frauenfelder, U. H., & Segui, J. (2004). Predictors of picture naming speed. *Behavior research methods, instruments, & computers*, *36*, 140–155.

Arndt, J., & Reder, L. M. (2002). Word Frequency and Receiver Operating Characteristic Curves in Recognition Memory: Evidence for a Dual-Process Interpretation. *Journal of experimental psychology. Learning, memory, and cognition*, *28*(5), 830–842.

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2019). Consistency and Variability in Children's Word Learning Across Languages. *Open Mind: Discoveries in Cognitive Science*, *3*, 52–67.

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2. *Journal of Experimental Psychology: Human Perception and Performance*, *42*, 441–458. (Place: US Publisher: American Psychological Association)

Calude, A. S., & Pagel, M. (2011). How do we use language? Shared patterns in the frequency of word use across 17 world languages. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1567), 1101–1107.

Calude, A. S., & Pagel, M. (2014). *Frequency of use and basic vocabulary.* John Benjamins Publishing Company.

Clark, S. E. (1992). Word frequency effects in associative and item recognition. *Memory & Cognition*, *20*(3), 231–243.

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The "small world of words" english word association norms for over 12,000 cue words. *Behavior research methods*, *51*, 987–1006.

Fellbaum, C. (1998). *WordNet: An electronic lexical database and some of its applications.* MIT press Cambridge.

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., Augustinova, M., & Pallier, C. (2010). The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods*, *42*(2), 488–496.

Gregg, V. (1976). Word frequency, recognition and recall. In *Recall and recognition* (pp. x, 275–x, 275). Oxford, England: John Wiley & Sons.

Grieve, J., Nini, A., & Guo, D. (2017). Analyzing lexical emergence in Modern American English online1. *English Language & Linguistics*, *21*(1), 99–127.

Halgren, E., & Smith, M. (1987). Cognitive evoked potentials as modulatory processes in human memory formation and retrieval. *Human neurobiology*, *6*(2), 129–139.

Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009a). Longitudinal analysis of early semantic networks: Preferential attachment or preferential acquisition? *Psychological science*, *20*(6), 729–739.

Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009b). Longitudinal Analysis of Early Semantic Networks: Preferential Attachment or Preferential Acquisition? *Psychological Science*, *20*(6), 729–739. (Publisher: SAGE Publications Inc)

Jescheniak, J. D., & Levelt, W. J. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of experimental psychology: learning, Memory, and cognition*, *20*(4), 824.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: from eye fixations to comprehension. *Psychological review*, *87*(4), 329.

Karjus, A., Blythe, R. A., Kirby, S., & Smith, K. (2020). Communicative need modulates competition in language change. *arXiv preprint arXiv:2006.09277*.

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, *44*(1), 287–304.

Laurino, J., De Deyne, S., Cabana, Á., & Kaczer, L. (2023). The pandemic in words: Tracking fast semantic changes via a large-scale word association task. *Open Mind*, *7*, 221–239.

Li, Y., Breithaupt, F., Hills, T., Lin, Z., Chen, Y., Siew, C. S., & Hertwig, R. (2024). How cognitive selection affects language change. *Proceedings of the National Academy of Sciences of the United States of America*, *121*(1).

Liu, Q., De Deyne, S., Jiang, X., & Lupyan, G. (2023). Understanding the frequency of a word by its associates: A network perspective. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 45).

Meier, B., Rey-Mermet, A., Rothen, N., & Graf, P. (2013). Recognition memory across the lifespan: the impact of word frequency and study-test interval on estimates of familiarity and recollection. *Frontiers in Psychology*, *4*.

Oldfield, R. C., & Wingfield, A. (1965). Response latencies in naming objects. *Quarterly Journal of Experimental Psychology*, *17*(4), 273–281.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive psychology*, *8*(3), 382–439.

Speer, R. (2022, September). *rspeer/wordfreq: v3.0.* Zenodo.

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive science*, *29*(1), 41–78.

Wang, S., & Bond, F. (2013). Building the Chinese open Wordnet (COW): Starting from core synsets. In *Proceedings of the 11th workshop on Asian language resources* (pp. 10–18). Nagoya, Japan: Asian Federation of Natural Language Processing.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, *46*, 441–517. (Place: Netherlands Publisher: Elsevier Science)

A cautionary note on sociodemographic predictors of linguistic complexity: different measures and different analyses lead to different conclusions

Gary Lupyan[*1] and Limor Raviv[2,3]

[*]Corresponding Author: lupyan@wisc.edu
[1]Department of Psychology, University of Wisconsin-Madison, Madison, WI, USA
[2]LEADS group, Max Planck Institute for Psycholinguistics, Nijmegen, NL
[3]cSCAN, University of Glasgow, Glasgow, UK

The question of why languages differ in the ways they do has been of longstanding interest in the fields of language evolution and language diversity. In 2010, Lupyan and Dale took advantage of the recently digitized World Atlas of Language Structures (WALS; Haspelmath et al., 2008) to test a hypothesis variously articulated by Trudgill (2002), Wray and Grace (2007) and McWhorter (2007). The broader claim was that some linguistic differences may arise from languages "adapting" to different sociodemographic environments. The more specific claim was that languages with histories of use by larger and more diverse speaking populations will tend to lack features such as complex agreement and inflectional systems that are thought to be difficult for nonnative learners (i.e., outsiders) to master. In line with this hypothesis, Lupyan & Dale, 2010 found that languages spoken by more people and spread over a larger area were more likely to use more lexical rather than inflectional means of communicating various information such as aspect, evidentiality, and possibility, and to systematically have fewer grammatical distinctions in e.g., types of possession, remoteness distinctions in tense, and grammatical encoding of space in demonstratives.

Numerous correlational studies have since confirmed this general trend (e.g., Bentz & Winter, 2013; Bentz et al., 2015; Nettle, 2012), but new analyses come with new caveats, e.g, some finding that the proportion of L2 speakers matters (Sinnemäki & Garbo, 2018), while others finding that it does not (Koplenig, 2019). Importantly, the link between group size and language structures has also begun to be experimentally tested, with studies finding that larger groups produce more systematic and compositional structure (Raviv et al., 2019, 2020). At the same time, developmental studies have been finding some evidence that children learn better from more redundant (i.e., complex) language input (Tal & Arnon, 2022; Portelance et al., 2023), helping to explain why languages may end up with such high levels of redundancy in the first place.

Recently, Shcherbakova et al. (2023) conducted a meticulous analysis of the link between grammatical complexity and sociodemographic factors and came to a very different conclusion, finding either no link or a *positive* relationship between complexity and population size. Their conclusion–"societies of strangers do not speak less complex languages"–squarely contradicts the earlier results.

We conducted a reanalysis to better understand what accounted for the qualitative difference between earlier work and Shcherbakova et al.'s (2023) results. Compared to past work, Shcherbakova et al. used more complex areal and phylogenetic models to better control for non-independence of languages. But what is so puzzling is that their analysis failed to find a negative association between population size and complexity even in the raw data, *prior* to the areal and phylogenetic controls suggesting that the difference in phylogenetic controls was the main source of the discrepancy. Neither were the sociodemographic predictors since these matched earlier work. This leaves two main sources of difference: which languages were included in the sample, and how grammatical differences/complexity was quantified.

Shcherbakova et al.'s analysis used the newly available Grambank database (Skirgård et al., 2023). Despite including 1314 languages, the new sample only partially overlaps with the earlier WALS-based sample. For example, Shcherbakova et al's data includes long-extinct languages such as Ancient Hebrew and Ancient Greek (with 0 speakers) while omitting some large languages such as German, Spanish, Bengali, and Gujarati (which are included in WALS). Because very small-population and very large-population languages have larger influence on the regression models predicting complexity from (log-transformed) population size, even small differences in which languages are included can lead to large differences in results.

In WALS, grammatical features are coded using ordinal, nominal, and binary schemes while Grambank codes all features as binary. The binary coding simplifies analysis, but can radically change the relative weighing of variables, e.g., what was previously one variable (number of cases) becomes *n* variables (has ablative, has locative, etc.). In our re-analysis, we found that the features associated with smaller populations tended to be those previously described, e.g., more complex possessive markings (GB058, GB059), demonstratives (GB036), and distinctions in clusivity (GB028) and pronouns (GB0310). The features found to be associated *positively* with population such as politeness (GB415), diminutives (GB315), and passive markers (GB147) were not included in earlier analyses, helping to further explain the discrepancies in the results.

Different language databases, different measures, and different analyses can yield substantially different conclusions about the relationship between sociodemographic features and linguistic complexity. Our re-analysis suggests that Shcherbakova et al's (2023) does not supersede earlier results, but the broader coverage offered by Grambank combined with the more powerful areal and phylogenetic controls can be used to gain further insights into which aspects of language are and are not shaped by sociodemographic factors.

Acknowledgements

We thank Olena Shcherbakova for sharing the data and analyses with us.

References

Bentz, C., Verkerk, A., Kiela, D., Hill, F., & Buttery, P. (2015). Adaptive Communication: Languages with More Non-Native Speakers Tend to Have Fewer Word Forms. *PloS One*, *10*(6), e0128254.

Bentz, C., & Winter, B. (2013). Languages with More Second Language Learners Tend to Lose Nominal Case. *Language Dynamics and Change*, *3*(1), 1–27.

Haspelmath, M., Dryer, M., Gil, D., & Comrie, B. (2008). *The world atlas of language structures online.* Munich: Max Planck Digital Library.

Koplenig, A. (2019). Language structure is influenced by the number of speakers but seemingly not by the proportion of non-native speakers. *Royal Society Open Science*, *6*(2), 181274.

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PLOS ONE*, *5*(1), e8559.

Mcwhorter, J. (2007). *Language Interrupted: Signs of Non-Native Acquisition in Standard Language Grammars.* Oxford University Press, USA.

Nettle, D. (2012). Social scale and structural complexity in human languages. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1597), 1829–1836.

Portelance, E., Duan, Y., Frank, M. C., & Lupyan, G. (2023). Predicting Age of Acquisition for Children's Early Vocabulary in Five Languages Using Language Model Surprisal. *Cognitive Science*, *47*(9), e13334.

Raviv, L., Meyer, A., & Lev-Ari, S. (2019). Compositional structure can emerge without generational transmission. *Cognition*, *182*, 151–164.

Raviv, L., Meyer, A., & Lev-Ari, S. (2020). The Role of Social Network Structure in the Emergence of Linguistic Structure. *Cognitive Science*, *44*(8), e12876.

Shcherbakova, O., Michaelis, S. M., Haynie, H. J., Passmore, S., Gast, V., Gray, R. D., Greenhill, S. J., Blasi, D. E., & Skirgård, H. (2023). Societies of strangers do not speak less complex languages. *Science Advances*, *9*(33), eadf7704.

Sinnemäki, K., & Garbo, F. D. (2018). Language Structures May Adapt to the Sociolinguistic Environment, but It Matters What and How You Count: A Typological Study of Verbal and Nominal Complexity. *Frontiers in Psychology*, *9*, 342569.

Skirgård, H., et al.. (2023). Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, *9*(16), eadg6175.

Tal, S., & Arnon, I. (2022). Redundancy can benefit learning: Evidence from word order and case marking. *Cognition*, *224*, 105055.

Trudgill, P. (2002). *Sociolinguistic Variation and Change.* Georgetown University Press.

Wray, A., & Grace, G. (2007). The consequences of talking to strangers : Evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua*, *117*(3), 543–578.

Seeing signs of morphology: form-meaning relations in British Sign Language morphology is iconic for hearing non-signers

Hannah Lutzenberger[1], Neil Fox[2], Heidi Proctor[2], and Adam Schembri[*2]

[*]Corresponding Author: a.schembri@bham.ac.uk
[1]Department of Communication and Cognition, Tilburg University, Tilburg, The Netherlands
[2]English Language and Linguistics, University of Birmingham, Birmingham, United Kingdom

Iconicity is a core property of spoken and signed languages (Perniss et al., 2010), and taps into shared human cognition. Evidence comes from sign languages: the meanings of some iconic signs can be guessed without prior linguistic knowledge (Ortega et al. 2019), although this is not always the case (e.g., Sevcikova Sehyr & Emmorey 2019), and it has been shown that language knowledge considerably affects the perception of iconicity (Occhino et al. 2017). Iconicity drives the emergence of phonology and the lexicon in emerging sign languages (Sandler et al. 2011); moreover, many sign languages show great cross-linguistic similarities in their lexicons due to overlap in iconic motivations and mappings (e.g. Currie et al. 2002) and recruit articulatory properties for iconic purposes (e.g. Börstell et al. 2016 for articulatory plurality). Nevertheless, the bulk of studies on the pervasiveness of iconicity in sign languages focus on the lexicon with anecdotal mentions that sign language morphology is also shaped by iconicity as evidenced by considerable cross-linguistic similarities (e.g. Aronoff et al. 2005). Initial evidence suggests that one morphological aspect, namely derivational movement modification between nouns and verbs in American Sign Language, is not caused by cognitive biases but by language-internal factors (Pyers & Emmorey 2022). In our study, we follow up on this study by testing the sensitivity to the iconicity in several different morphological modifications of British Sign Language (BSL), aiming to contribute further evidence whether iconicity in sign language morphology arises due to shared aspects of human cognition.

We tackle this by testing hearing non-signers on whether they correctly identify the meaning of morphological modifications in BSL, using a forced-choice guessing task on Prolific. We presented participants with pairs of signed stimuli: a citation form and a form with morphological modification. Morphological modifications included verb directionality (modification of movement path), verbal plural (sweep vs. repeated movement), aspect marking (fast vs. slow movement), and non-manual modification (puffed cheeks vs. tongue protrusion). We recruited 100 hearing British participants without any sign language knowledge through Prolific (Female: 50; mean age: 40 years). Participants were presented with the citation form video accompanied by a single lexical equivalent in English and then asked to pick one of four possible English translations to match the sign with a morphological modification.

Combining descriptive and inferential statistics, our data suggests that BSL morphology is iconic even without language knowledge (Figure 1); the accuracy of hearing non-signers in picking the correct response is significantly better than chance ($p < 0.001$ in an intercept model with random intercepts for participant and item). To further explore this effect within each morphological modification, we conducted four exploratory analyses on each subset of the data: except for verb directionality, response accuracy is significantly predicted by morphological modification (mixed effects modeling with random intercepts for item and random slopes for participant).

Together, these findings suggest that iconicity in sign language morphology is accessible without sign language knowledge due to shared human cognition. Showing that hearing non-signers can access iconicity in morphological structures similar to how they are able to use it on the lexical level emphasizes the resilience of iconicity in sign languages and highlights the importance of the core property of iconicity in language emergence and evolution.

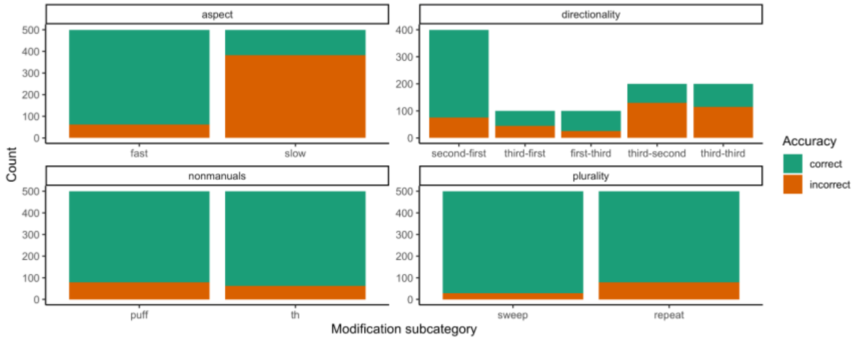


Figure 1: Overview of results.

References

Aronoff, M., Meir, I., & Sandler, W. (2005). The paradox of sign language morphology. *Language*, *81*(2), 301–344. https://doi.org/10.1353/lan.2005.0043

Börstell, C., Lepic, R., & Belsitzman, G. (2016). Articulatory plurality is a property of lexical plurals in sign language. *Lingvisticae Investigationes*, *39*(2), 391–407. https://doi.org/10.1075/li.39.2.10bor

Currie, A.-M., Meier, R., & Walters, K. (2002). A cross-linguistic examination of the lexicon of four signed languages. In R. P. Meier, K. Cormier, & D. Quinto-Pozos (Eds.), *Modality and structure in signed and spoken languages* (1st ed., pp. 224–236). Cambridge University Press. https://doi.org/10.1017/CBO9780511486777.011

Occhino, C., Anible, B., Wilkinson, E., & 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.16.1.04occ

Ortega, G., Schiefner, A., & Özyürek, A. (2019). Hearing non-signers use their gestures to predict iconic form-meaning mappings at first exposure to signs. *Cognition*, *191*, 103996. https://doi.org/10.1016/j.cognition.2019.06.008

Perniss, P., Thompson, R. L., & Vigliocco, G. (2010). Iconicity as a general property of language: Evidence from spoken and signed languages. *Frontiers in Psychology*, *1*, 1–15. https://doi.org/10.3389/fpsyg.2010.00227

Pyers, J. E., & Emmorey, K. (2022). The iconic motivation for the morphophonological distinction between noun–verb pairs in American Sign Language does not reflect common human construals of objects and actions. *Language and Cognition*, 1–23. https://doi.org/10.1017/langcog.2022.20

Sandler, W., Aronoff, M., Meir, I., & 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

Sevcikova Sehyr, Z. & Emmorey, E. (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 https://doi:10.1017/langcog.2019.18

Reconstructing a Protolanguage

Erkki Luuk[*], John Ioannis Stavroulakis[1,2]

[*]Corresponding Author: erkkil@gmail.com
[1]School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332, USA
[2]Department of Mathematics, Ariel University, Ariel 4076414, Israel

This paper documents an experimental reconstruction of a protolanguage. Combining research from diverse fields such as phonology, syntax and archaeology, we construct a plausible and expressive protolanguage vocabulary of 170 words, and provide examples of the protolanguage in use. We thus obtain a direct, reverse-engineered insight into the evolution from protolanguage to modern language.

1. Introduction

Protolanguage (PL) is the conjectured non-grammatical hominin language which was the precursor of modern fully syntactic languages (MLs), spoken roughly in the timeframe of 2 – 0.2 mya. The alternative, a Chomskyan theory of an abrupt emergence of ML from scratch ca. 70 – 100 kya (e.g. Bolhuis et al., 2014) does not pan out well. While a capacity for language can emerge abruptly (e.g. as a result of a mutation), the emergence of a ML implies a long and piecemeal process of inventing and tinkering with a large number of signs. Thus, the complexity of MLs implies a prehistoric existence of a PL. However, a PL is not directly attested and cannot be reconstructed by methods of historical linguistics, in contrast to, e.g. the Proto-World, the last common ancestor of all MLs (for the latter, see e.g. Starostin, 2019; http://ehl.santafe.edu/intro1.htm). Likewise, it is difficult to estimate the number of PLs that have existed[1] or the possible genealogical relationships between them — although one of them had to be the first chronologically.

Since the investigations of Swadesh (1972), research in various fields including phonetics, ideophones, historical linguistics, grammar, archeology, and medicine has shed much light on the evolution from PL to ML. The present paper combines such diverse insights, attempting a complete reconstruction of a plausible PL vocabulary (with poetic examples, see Supplementary Materials: bit.ly/46Kjf6z).

[1] See Johansson (2021) for some results in this direction.

Despite being a thought experiment, the reconstruction enables one to study rich examples of PL *in vivo*. We thus arrive at experimental, hands-on results on the evolution of language and the relations between diverse investigations of PL.

2. Phonology

The Frame/Content Theory of MacNeilage and Davis is one of the first to provide a basis for understanding the origin of the syllable and the evolution of phonemic inventories from PL to ML (see e.g. MacNeilage, 1994, 1998; Davis & Zajdó, 2008). This approach considers the CV syllable cycle to constitute the basis of all speech, drawing on similarities with the mandibular cycle. It considers the evolution from baby "babbling" to adult speech as parallel to the evolution of phonemic inventories from PL to ML, with the phonemes that are spoken earlier occurring more frequently in the world's languages (also noting the correlation between certain vowels and consonants, e.g. velars and back vowels).

The reconstruction of PL phonemes of MacNeilage (1994) is one of the main inspirations of the present paper. The authors improve upon this by extrapolating to a simpler phonemic inventory to reach a more distant past. Considering its gradual evolution from nonlinguistic calls (cf. MacNeilage, 1994), one can assume a limited phoneme inventory for PL. Taking into account the phonological universals as well as the phonemics of "babbling" (Stefanuto & Vallée, 1999; Hyman, 2007), there emerges a minimal inventory of six consonants and three vowels as the limit of what one can extrapolate from phonological typology. The most plausible such inventory would be akin to /p, t, k, b, d, g, i, a, u/ (see Stefanuto & Vallée, 1999; Hyman, 2007: 345-359). Amongst 3-vowel systems, /i, a, u/ is the most common, and the Rotokas language constitutes a foremost example of 6-consonant systems (Firchow & Firchow, 1969; Hyman, 2007: 349-351). The phonemic inventory used in this paper is a variation on the Rotokas theme, /p, t, k, m, l, ɣ, i, a, u/ (/m/, /l/, /ɣ/ are allophones of /b/, /d/, /g/, respectively). The letter symbols could, in principle, stand for any 6-consonant and 3-vowel system, and in any case, one expects much allophonic variation in languages with a limited inventory. To reiterate, the PL constructed in this paper is a *possible* PL. We make no claim as to the occurrence/attestation of these particular phonemic and lexical inventories beyond their possibility and partial plausibility, largely due to the enormous number of PLs (Johansson, 2021).

At the other end of the spectrum, one considers non-human primate calls, and the speech capacities of early hominin physiology during the evolution from PL to ML. Neanderthals likely had speech capacities similar to modern *H. sapiens* (see the discussion in Lieberman & Crelin, 1971; Albanese, 1994; Lieberman, 2002; Boë et al., 2007; Barney et al., 2012; Conde-Valverde et al., 2021). Given the blurry distinction and lineage between *Australopithecus* and *Homo* (Bruner & Baudet, 2023; Herries et al., 2020; Kimbel & Villmoare, 2026), the linguistic capacities of early members of *Homo* remain uncertain, though see MacLarnon & Hewitt (1999), Meyer et al. (2006) and Tobias (1998). On gauging the phonetics and origin of protospeech by studying other primates and the evolution of "speech organs" in general, the reader is referred to Boë et al. (2017).

The correlation between the evolution of phonemic inventories and genetic evolution is certainly an interesting open problem. It is clear that, given specific constraints on the syllable/word structure, the phonemic inventory is directly related to the size of the lexicon. Martinet (1957, 1960a, 1960b) and Hockett & Hockett (1960) independently introduced the concept of *duality of patterning* or *double articulation*, the property of meaningful words being comprised of combinations of meaningless phonemes. Hockett & Hockett (1960) considered duality to be the final step in the evolution of language from primate calls, satisfying the need for clarity in a large vocabulary. For a detailed study of duality and its other possible motivations such as predictability/learnability and conventionalization, see e.g. Del Giudice et al. (2010), Jackendoff (1999), Nowak & Krakauer (1999), Pinker & Jackendoff (2005), Sandler et al. (2011), Verhoef et al. (2014) and the references therein.

An interesting question is whether duality was a characteristic of all PLs. The study of Sandler et al. (2011) indicates that duality, even if not *a priori* necessary, becomes inevitable over the course of time (especially in vocal as opposed to gestural modality). In the absence of a grammar governing morphological transformations, the definition of duality for PL may not be straightforward, as it presupposes a functional distinction between *phonemes* and *words*. It is plausible, however, that most PLs had a far greater number of words than phonemes, and this statistical distinction would be equivalent to duality. It would be interesting to investigate the close relationship between duality and the size of the lexicon with e.g. a certain syllable structure (cf. Jackendoff, 1999, and the references therein). Concerning the PL reconstructed here, the 6 consonant, 3 vowel inventory is a plausible extrapolation at the limit of attested inventories, while still allowing for sufficient lexical freedom. We finally remark that significant evolution of human sound systems has occurred even during the last few ky — for example, post-neolithic changes in diet favored the occurrence of labiodentals such as /f/, /v/ (Blasi et al., 2019).

3. Sound-Symbolism

The Saussurean arbitrariness of the sign has been called into question by recent research on sound-symbolism as a universal tendency of language (see Voeltz & Kilian-Hatz, 2001; Lockwood & Dingemanse, 2015; Westbury et al., 2018). Moreover, the seminal results of Swadesh (1972) agree with the current state of the art. Remarkably, Swadesh (1972), Lockwood & Dingemanse (2015) and Blasi et al. (2016) are mutually compatible theories of sound-symbolism.

In constructing the bulk of the vocabulary of our PL, we have attempted to use the most popular combinations of sound and meaning found in the ASJP database (Wichmann et al., 2020). The research on sound-symbolism and ideophones confirms that such popular (statistically relevant) combinations are not random statistical outliers, but often correlate with sound-symbolism. It is also possible that universal principles other than sound-symbolism contribute to such tendencies. It is an open question what those principles would be, and whether their identification as "sound-symbolism" is a matter or definition or not.

Studies on sound-symbolism (such as Swadesh, 1972; Lockwood & Dingemanse, 2015; Westbury et al., 2018) were particularly useful in creating a plausible PL vocabulary, both as starting points and as guiding principles. For example, *muki, miki, lila*, "man", "boy", "woman", are partly inspired by *impicic, alyel* of Westbury et al. (2018, Table 9). In order to create a PL as simple and archaic as possible, the etyma were constructed to be short and regular: all syllables are CV, in accordance with the Frame/Content Theory, and words contain at most two syllables.

As a historical note, we remark that while the "Global Etymologies" of Bengtson & Ruhlen (1994) are rightfully criticized by many for inconsistencies such as allowing for the construction of fallacious etyma (e.g. Campbell, 2008), several such etyma do describe sound-symbolism. The intuitions of Bengtson and Ruhlen pinpointed not correct etymologies but certain sound-symbolic tendencies of language. Cabrera (2012) shows that two of the "global etyma" can be derived from the sound-symbolic theories of Swadesh (1972) and well-known archeological evidence of symbolism. Other false "global etyma" can be constructed almost directly from the sound-symbolic tendencies documented in Blasi et al. (2016: 10820). E.g. the etyma *buka*, *čunga*, resemble very closely the /o, u, p, k, q/ and /u, n/ phonemes predicted by Blasi et al. (2016: 10820), for the meanings "knee" and "nose", respectively. While Bengtson & Ruhlen (1994) set out to find global etyma, they didn't, but found sound-symbolism instead.

4. The Origins of Syntax and Morphology

PL is, by definition, a hominin language without syntax and morphology, which was the precursor of modern (fully) syntactic languages. An alternative, weaker definition would be a language without grammatical morphemes (cases, adpositions, articles, etc.). A word order rule would automatically result in different grammatical categories for the words, inducing a syntax and grammar even in the absence of such morphemes (Luuk, 2012). Given the number of extinct hominin species and the possibility that *H. habilis* was a speaker (see Tobias, 1998, and the tentative timeframe in Luuk, 2018), one is inclined to assume a staggering diversity of PLs over a very long period (perhaps even one million years). The diversity would be due to the great time span, the number of distinct hominin species, and the difficulty of discerning a stable stage in the evolution of PL (from the vantage point of ML).

The essential continuity between embedding and non-embedding, modal markers and non-modal words, etc. (cf. Evans & Levinson, 2009, 2.2.5; Luuk, 2013) would make the evolution from PL to ML gradual. Discussions about e.g.

the status of embedding in Pirahã (e.g. Everett, 2009; Nevins et al., 2009) underline such ambiguities.

The evolution from non-human primate calls to PL was even more gradual and remains harder to track. While primates are capable of producing, communicating and combining a variety of signs (Engesser et al., 2016; Leroux & Townsend, 2020) and possess a rudimentary understanding of message structure, there remains a stark gap between human and non-human primate capacities (Wang et al., 2015; Jiang et al., 2018) — a difference in quality of signal processing, or a difference in quantity and speed large enough to be *de facto* qualitative.

Either a fractionation or concatenation of the existing signs could explain the evolution of language. A closely related question is whether the advent of morphology preceded that of syntax or vice versa. Fractionation (see e.g. Wray, 1998) might explain the origin of nonconcatenative morphology, as well as its seeming “persistence” and universality (Dubé, 2011), even in pidgins and creoles (Plag, 2006). However, fractionation is dubious and statistically unfeasible (Bickerton, 2003; Johansson, 2008). Further, (nonconcatenative) morphology seems to owe its “persistence” to the ease with which it is learned by children (Ratcliffe, 2007) and its origin to concatenative grammar (Svenonius & Bye, 2012). Thus, it is assumed that grammar traces its origins to the concatenation of signs and syntax (e.g. Jackendoff, 1999; Luuk, 2013; Nowak et al., 2001).

5. Vocabulary

The “Semantic Universals” project of Goddard and Wierzbicka is a philosophically ambitious one, including the analysis of — presumably universal — semantic units from which all other meanings can be constructed (see Goddard & Wierzbicka, 1994; Goddard, 2001, and the references therein). It has been successful in revealing several semantic units that are reliably lexicalized across the world’s languages, improving on previous work such as Swadesh’s lexicostatistical list (Swadesh 1972: 283-284). The semantic universals list (Goddard & Wierzbicka, 1994, 2.2; Wierzbicka, 1996; Goddard, 2001, Table 3) and the discussion in Goddard & Wierzbicka (1994, 2.2) and Goddard (2001, 2) were a useful basis in forming a PL vocabulary. Several modal etyma were modified or replaced with non-grammatical equivalents, implementing ideas from Goddard (2001, 2) and Evans & Levinson (2009, 2.2.5). For example, instead of “or” and “if”, “may(be)” was used, in accordance with the Guugu Yimidhirr language (see Haviland, 1979).

The semantic universals formed the scaffolding around which the senses of each word were organized. As the PL has a limited lexicon, it makes extensive use of *colexification*, association of different senses with the same lexical form. Since the seminal paper of François (2008), several studies have shed light on its motivation and function (Di Natale et al., 2011; Karjus et al., 2021; Brochhagen

& Boleda, 2022). Colexification boosts efficiency of communication by expressing related concepts with the same word when there is no risk of ambiguity. It also has interesting relations to the evolution of metaphor and highlights the simultaneous pressures for simplicity and clarity/information that influence the lexicon. There exist several cross-linguistic databases of colexification, the most extensive of which is the CLICS colexification database (Rzymski et al., 2019, https://clics.clld.org). The reader is referred to the discussion in Brochhagen & Boleda (2022: 2) for more information on such databases. CLICS was consulted throughout the formation of the PL vocabulary, guiding the association of senses/meanings to the etyma.

Another valuable insight into languages with limited vocabulary was Toki Pona, the philosophical language created by Sonja Lang (Lang, 2014; https://tokipona.org). Similarly to PL, Toki Pona operates with minimal grammar, vocabulary, and phoneme inventory, achieving in philosophy what the present paper does in linguistics.

Since PL had (by definition) no grammatically distinct word classes, a universal flexible word class (e.g. Luuk, 2010; Rijkhoff & van Lier, 2013) has been used extensively. For example, *kuka* means both “crawl” and “reptile”, *γuγi* “slither” and “snake”, *kipa* “cleave” and (the tool) “cleaver”, etc. In the absence of distinct word classes, predicates and arguments were identified by their semantics and pragmatic context. E.g. *muma kapa*, literally “many/much oar/paddle”, either “many oars” or “paddle much”, can be resolved to the second meaning in the context of two (*pali*) persons going (i.e. sailing, *ma*) at sea (*γuma*).

Special attention was paid to accommodating words for the cultural artifacts of the Lower Paleolithic (spears, cupules, pigment, etc.), as well as for the relevant social (gang, friend, wife, foreign, etc.), kinship (father, mother, son, daughter, etc.), natural (tree, stone, hunt, etc.) and body part (leg, hand, belly, head, etc.) categories. The result of the work was a compact yet rich vocabulary of 170 words, which allows to convey information on diverse topics from subsistence to mythology, from natural to technological and conceptual to bodily domains, and even compose poems (see Supplementary Materials: bit.ly/46Kjf6z).

Acknowledgements

The second author acknowledges support of Ariel University during the initial stages of this research.

References

Albanese, J. (1994). Neanderthal speech. *The University of Western Ontario Journal of Anthropology*, *1*(1).

Barney, A., Martelli, S., Serrurier, A., & Steele, J. (2012). Articulatory capacity of Neanderthals, a very recent and human-like fossil hominin. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1585), 88-102.

Bengtson, J. D., & Ruhlen, M. (1994). Global etymologies. *On the origin of languages: studies in linguistic taxonomy*, 277-336.

Bickerton, D. (2003). Symbol and structure: a comprehensive framework for language evolution. *Studies in the Evolution of Language*, *3*, 77-93.

Blasi, D. E., Moran, S., Moisik, S. R., Widmer, P., Dediu, D., & Bickel, B. (2019). Human sound systems are shaped by post-Neolithic changes in bite configuration. *Science*, *363*(6432).

Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. F., & Christiansen, M. H. (2016). Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences*, *113*(39), 10818-10823.

Boë, L. J., Fagot, J., Perrier, P., & Schwartz, J. L. (2017). *Origins of human language: Continuities and discontinuities with nonhuman primates*. Peter Lang International Academic Publishers.

Boë, L. J., Heim, J. L., Honda, K., Maeda, S., Badin, P., & Abry, C. (2007). The vocal tract of newborn humans and Neanderthals: Acoustic capabilities and consequences for the debate on the origin of language. A reply to Lieberman (2007a). *Journal of Phonetics*, *35*(4), 564-581.

Bolhuis, J. J., Tattersall, I., Chomsky, N., Berwick, R. C. (2014) How Could Language Have Evolved? *PLOS Biology* 12(8): e1001934.

Brochhagen, T., & Boleda, G. (2022). When do languages use the same word for different meanings? The Goldilocks principle in colexification. *Cognition*, *226*, 105179.

Bruner, E., & Beaudet, A. (2023). The brain of Homo habilis: Three decades of paleoneurology. *Journal of Human Evolution*, *174*, 103281.

Cabrera, J. C. M. (2012). The role of sound symbolism in protolanguage: Some linguistic and archaeological speculations. *Theoria et Historia Scientiarum*, *9*, 115-130.

Campbell, L. (2008). *What can we learn about the earliest human language by comparing languages known today?,* In Bernard Laks, Serge Cleuziou, Jean-Paul Demoule & Pierre Encrevé (eds.), *Origins and evolution of language: Approaches, models, paradigms* (pp. 79-111). London: Equinox Publishing Limited.

Conde-Valverde, M., Martínez, I., Quam, R. M., Rosa, M., Velez, A. D., Lorenzo, C., Castro, J. M. B. (de), Carbonell, E., & Arsuaga, J. L. (2021). Neanderthals and Homo sapiens had similar auditory and speech capacities. *Nature Ecology & Evolution*, *5*(5), 609-615.

Davis, B. L., & Zajdó, K. (eds.) (2008). *The Syllable in Speech Production*. New York, London, Taylor & Francis Group.

Del Giudice, A., Kirby, S., & Padden, C. (2010). Recreating duality of patterning in the laboratory: a new experimental paradigm for studying emergence of sublexical structure. In *The Evolution of Language* (pp. 399-400).

Di Natale, A., Pellert, M., & Garcia, D. (2021). Colexification networks encode affective meaning. *Affective Science*, *2*(2), 99-111.

Dubé, J. (2011). Reconsidering the "isolating protolanguage hypothesis" in the evolution of morphology. In *Annual Meeting of the Berkeley Linguistics Society* (Vol. 37, No. 1, pp. 76-90).

Engesser, S., Ridley, A. R., & Townsend, S. W. (2016). Meaningful call combinations and compositional processing in the southern pied babbler. *Proceedings of the National Academy of Sciences*, *113*(21), 5976-5981.

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, *32*(5), 429-448.

Everett, D. L. (2009). Pirahã culture and grammar: A response to some criticisms. *Language*, 405-442.

Firchow, I., & Firchow, J. (1969). An abbreviated phoneme inventory. *Anthropological Linguistics*, 271-276.

François, A. (2008). Semantic maps and the typology of colexification: Intertwining polysemous networks across languages. In *Studies in language companion series* (pp. 163–215). John Benjamins Publishing Company.

Goddard, C. (2001). Lexico-semantic universals: A critical overview. *Linguistic typology*, *5*(1), 1-65.

Goddard, C., & Wierzbicka, A. (Eds.). (1994). Semantic and lexical universals: Theory and empirical findings.

Herries, A. I.R., et al. (2020). Contemporaneity of australopithecus, paranthropus, and early homo erectus in South Africa. *Science*, 368.6486: eaaw7293.

Hockett, C. F., & Hockett, C. D. (1960). The origin of speech. *Scientific American*, *203*(3), 88-97.

Hyman, L. M. (2008). Universals in phonology. *The Linguistic Review*, 25, 83–137.

Jackendoff, R. (1999). Possible stages in the evolution of the language capacity. *Trends in Cognitive Sciences*, *3*(7), 272-279.

Jiang, X., Long, T., Cao, W., Li, J., Dehaene, S., & Wang, L. (2018). Production of supra-regular spatial sequences by macaque monkeys. *Current Biology*, *28*(12), 1851-1859.

Johansson, S. (2008). Seeking compositionality in holistic proto-language without substructure – do counterexamples overwhelm the fractionation process? In: A. D. M. Smith, K. Smith & R. Ferrer i Cancho (eds.), *The Evolution of Language: Proceedings of the 7th International Conference on the Evolution of Language* (EVOLANG7), pp. 171-178. Singapore, New Jersey, etc.: World Scientific.

Johansson, S. (2021). How many protolanguages were there?. *Ways to Protolanguage* (Protolang 7). https://blogs.phil.hhu.de/protolang7/how-many-protolanguages-were-there

Karjus, A., Blythe, R. A., Kirby, S., Wang, T., & Smith, K. (2021). Conceptual similarity and communicative need shape colexification: An experimental study. *Cognitive Science*, *45*(9), e13035.

Kimbel, W. H., & Villmoare, B. (2016). From Australopithecus to Homo: the transition that wasn't. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1698), 20150248.

Lang, S. (2014). *Toki Pona: The Language of Good*. CreateSpace.

Leroux, M., & Townsend, S. W. (2020). Call combinations in great apes and the evolution of syntax. *Animal Behavior and Cognition*, *7*(2), 131-139.

Lieberman, P. (2002). On the nature and evolution of the neural bases of human language. *Yearbook of Physical Anthropology*, 45, 36–62

Lieberman, P., & Crelin, E. S. (1971). On the Speech of Neanderthal Man. *Linguistic Inquiry*, *2*(2), 203–222.

Lockwood, G., & Dingemanse, M. (2015). Iconicity in the lab: A review of behavioral, developmental, and neuroimaging research into sound-symbolism. *Frontiers in Psychology*, *6*, 1246.

Luuk, E. (2010). Nouns, verbs and flexibles: implications for typologies of word classes. *Language Sciences, 32(*3), 349-365.

Luuk, E. (2012). The origins of linguistic predicate/argument structure. In: T. C. Scott-Phillips, M. Tamariz, E. A. Cartmill & J. R. Hurford (eds.), *The Evolution of Language. Proceedings of the 9th International Conference* (EVOLANG9), pp. 204-211. Singapore, New Jersey, etc.: World Scientific.

Luuk, E. (2013). The structure and evolution of symbol. *New Ideas in Psychology*, *31*(2), 87-97.

Luuk, E. (2018). Symbol and its evolution. In Cuskley, C., Flaherty, M., Little, H., McCrohon, L., Ravignani, A. & Verhoef, T. (Eds.): *The Evolution of Language: Proceedings of the 12th International Conference* (EVOLANGXII).

MacLarnon, A. M., & Hewitt, G. P. (1999). The evolution of human speech: The role of enhanced breathing control. *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists*, *109*(3), 341-363.

MacNeilage, P. (1994). Prolegomena to a Theory of the Sound Pattern of the First Spoken Language. *Phonetica*, 51(1-3), 184-194.

MacNeilage, P. (1998). The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences*, 21(4), 499-511.

Martinet, A. (1957). Arbitraire linguistique et double articulation. *Cahiers Ferdinand de Saussure*, 15, 105–116.

Martinet, A. (1960a). Elements of a functional syntax. *Word*, *16*(1), 1-10.

Martinet, A. (1960b). *Éléments de linguistique générale*. Paris: Armand Collin.

Meyer, M. R., Lordkipanidze, D., & Vekua, A. (2006). The anatomical capacity for spoken language in Homo erectus. *Am. J. Phys. Anthropol.*, *129*, 130.

Nevins, A., Pesetsky, D., & Rodrigues, C. (2009). Pirahã exceptionality: A reassessment. *Language*, 355-404.

Nowak, M. A., & Krakauer, D. C. (1999). The evolution of language. *Proceedings of the National Academy of Sciences*, *96*(14), 8028-8033.

Nowak, M. A., Komarova, N. L., & Niyogi, P. (2001). Evolution of universal grammar. *Science*, *291*(5501), 114-118.

Plag, I. (2006). Morphology in pidgins and creoles. In K. Brown (ed.), *Encyclopedia of Language and Linguistics*, vol. 8, 304–308. Oxford: Elsevier.

Pinker, S., & Jackendoff, R. (2005). The faculty of language: what's special about it?. *Cognition*, *95*(2), 201-236.

Ratcliffe, R. (2007). Why do languages develop and maintain non-concatenative morphology?. In *Mediterranean Morphology Meetings* (Vol. 6, pp. 99-115).

Rijkhoff, J. and van Lier, E. (eds.). (2013). Flexible Word Classes: Typological studies of underspecified parts of speech. Oxford: Oxford University Press.

Rzymski, Christoph and Tresoldi, Tiago et al. (2019). The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies.

Sandler, W., Aronoff, M., Meir, I., & Padden, C. (2011). The gradual emergence of phonological form in a new language. *Natural language & linguistic theory*, *29*, 503-543.

Starostin, G. (2019). Macro-comparative linguistics in the 21st century: state of the art and perspectives. *Journal of Language Relationship*, 1-12.

Stefanuto, M., & Vallée, N. (1999). Consonant systems: From universal trends to ontogenesis. In *ICPhS, San Francisco*.

Swadesh, M. (1972). *The origin and diversification of language*. London: Routledge & Kegan Paul.

Svenonius, P., & Bye, P. (2012). Non-concatenative morphology as epiphenomenon. in J. Trommer (ed.), *The morphology and phonology of exponence* (Oxford Studies in Theoretical Linguistics 41). Oxford: Oxford University Press, 427-95.

Tobias, P. V. (1998). Evidence for the early beginnings of spoken language. *Cambridge Archaeological Journal*, 8(1), 72–78.

Verhoef, T., Kirby, S., & De Boer, B. (2014). Emergence of combinatorial structure and economy through iterated learning with continuous acoustic signals. *Journal of Phonetics*, *43*, 57-68.

Voeltz, F. E., & Kilian-Hatz, C. (Eds.). (2001). *Ideophones* (Vol. 44). John Benjamins Publishing.

Wang, L., Uhrig, L., Jarraya, B., & Dehaene, S. (2015). Representation of numerical and sequential patterns in macaque and human brains. *Current Biology*, *25*(15), 1966-1974.

Westbury, C., Hollis, G., Sidhu, D. M., & Pexman, P. M. (2018). Weighing up the evidence for sound symbolism: Distributional properties predict cue strength. *Journal of Memory and Language*, *99*, 122-150.

Wichmann, Søren, Eric W. Holman, and Cecil H. Brown (eds.). 2020. The ASJP Database (version 19).

Wray, A. (1998). Protolanguage as a holistic system for social interaction. *Language & Communication*, *18*(1), 47-67.

Do Dogs Really Get the Point?

Heidi Lyn[1,2*], Joclyn Villegas[2], Christopher Bass[1], Steven Baker[1], Katie West[1], Lindsey Johnson[1]
[*]Corresponding Author: hlyn@southalabama.edu
1 Department of Psychology 2 School of Marine Sciences, University of South Alabama, Mobile, AL, USA

Over the past two decades, researchers have discussed the seeming anomaly of dogs' ability to follow human pointing gestures more successfully than great apes or wild canids and have suggested these findings were due to domestication instilling a biological capacity to communicate with humans (e.g. Hare at al., 2010). Other researchers, however, found that some wild species do follow points at similar rates to domesticated dogs, challenging initial biological theories in favor of human interaction or learning theories (Udell Dorey, & Wynne, 2008). Similarly, some studies have found that variations in point type and the presence or absence of ostensive cues can greatly affect the success of point following (Lyn, et al., 2021). These findings call into question the communicative nature of canine point following and its potential connection to language evolution. To investigate these and other questions, we initiated a meta-analysis of the over 20 years of dog pointing studies.

A total of 146 peer-reviewed articles were included in the initial analysis, collected from wide-ranging database search using the keywords: point, pointing, cues, gesture, communication, dog, canine, distal, proximal, ostension, human gesture, communicative, and object-choice task. This study focuses on a data set of 55 articles that included individual data for the dogs, comprising 1465 individual data points, of which 1173 were pet dogs, 113 were a wild canid species (wolf, fox, etc.), 94 were research dogs and 85 were shelter dogs. Dogs were relatively evenly split between males and females (649 males, 665 females, 151 unrecorded). Average age was 2.47 with over 1/3 of the dogs under 7 months.

Contrary to many earlier studies, our meta-analysis found no systematic differences between wild canid species and domesticated dogs, instead, the only group that was different to all other groups was shelter dogs, who performed

more poorly than every other group (F(3, 186) = 15.5, p<001, Games-Howell post-hoc test p<.003 for all comparisons with shelter dogs).
The dogs' known socialization history with humans did have an effect on pointing comprehension, but not for typical amounts of socialization (regular pet interactions), dogs only performed better if they were recorded to be in contact with humans constantly (F(3, 178)=5.63, p=.001). Dogs did seem to learn to follow points quickly; percentage of points correct was positively correlated with the number of trials the dog completed (r=.12, p<.001, maximum number of trials = 30). Notably, our data show that dogs react differently depending on the type of point on offer and the ostensive cues that are present (F(1, 1461) = 53.3, p<.001). Dogs follow ipsilateral (same side) points equally well with and without a gaze cue, but responses to contralateral points fall to chance levels without gaze cues.

Our results suggest that dogs don't comprehend points as a communicative mechanism, rather they likely learn to follow points primarily through associative mechanisms such as location of movement or distance between gesturing arm and object. When those cues are removed, as when the gesture crosses the researchers' bodies, the dogs fall to chance levels, unless another cue is added. Further, the domestication process does not seem to confer any benefits onto dogs that wild canids did not already have. Constant interaction with humans did seem to increase point following, suggesting a strong learned component. Future research must first and foremost avoid presumption of communicative mechanisms when interpreting dogs point following, and carefully analyze any suggested links to language evolution.

Acknowledgements

Heidi Lyn holds the Joan M. Sinnott Chair which is funded by the USA Foundation.

References

Hare, B., Rosati, A., Kaminski, J., Bräuer, J., Call, J., & Tomasello, M. (2010). The domestication hypothesis for dogs' skills with human communication: A response to Udell et al. (2008) and Wynne et al. (2008). Animal Behaviour, 79(2), e1–e6..

Lyn, H., Broadway, M., Jett, S. E., Samuelson, M. M., Christopher, J., & Chenkin, B. (2021). The effects of distance on pointing comprehension in shelter dogs. *Animal Cognition*, *24*, 855-865..

Udell, M. A. R., Dorey, N. R., & Wynne, C. D. L. (2008). Wolves outperform dogs in following human social cues. *Animal Behaviour*, *76*(6), 1767–1773.

A feedback-facilitated iterated learning experiment

Calen MacDonald*[1], Morten H. Christiansen[1,2]

*Corresponding Author: ccm262@cornell.edu
[1]Psychology Department, Cornell University, Ithaca, USA
[2]School of Communication and Culture, Aarhus University, Aarhus, Denmark

Many languages like Korean, Russian, and Turkish use flexible word orders to express the same meanings, often with different pragmatic emphasis. Yet, iterated learning experiments suggest that cultural transmission eliminates alternate patterns. How do flexible word orders persist across generations of learners?

Languages in iterated learning experiments tend to degenerate without pressures like horizontal transmission (Carr et al., 2017; Theisen-White et al., 2011) or filtering (Kirby et al., 2008). Feedback might have a similar influence. It improves learning in artificial languages (Jeuniaux et al., 2009; Monaghan et al., 2021) and the L2 (Carroll & Swain, 1993; Ellis et al., 2006). Like horizontal transmission, feedback models the sociolinguistic factors that predominately explain trade-offs between linguistic variables, such as word order acting as a cue for grammatical role in languages without case marking (Levshina, 2021). Therefore we consider that feedback may help maintain multiple alternate non-native word orders in a culturally transmitted artificial language.

Materials. Participants learned 24 sentences comprising a verb, subject, and case-marked object, corresponding each to one of 24 scenes. The initial language contained 12 VOS and 12 VSO sentences. All sentences used the same verb (*poox*) and object case marker (*-ma*). The language's eight nouns were English words, divided into sets for Exposure (*goat, cat, fox, horse*) and Generalization (*dog, sheep, pig, cow*).

Sentences described "who does what to whom" scenes picturing a subject animal and a red arrow (*poox*) pointing to an object animal. Because VSO and VOS sentences involve different pragmatic emphasis, a blue star marked the referent that is foregrounded by being mentioned first (Fig 1).

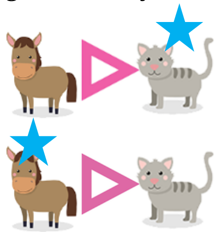


Fig 1. In the initial language: *Poox cat-ma horse* (top), *Poox horse cat-ma* (bottom)

Method. The experiment had three parts: Noun, Exposure, and Generalization. In each part, participants completed two types of trials: Comprehension and Production. Comprehension trials presented text for two seconds then a two alternative forced choice (2AFC) between images. Production trials presented images for two seconds then a three alternative forced choice (3AFC) between texts. The initial Noun part consisted of eight trials of both types and introduced participants to the experiment using familiar English nouns.

Next, the Exposure part contained six blocks of 24 trials of each type covering all 24 scene-sentence pairs. To correctly answer a 2AFC trial required learning the object case marker as one answer choice was the correct scene and the foil reversed that scene's subject and object. A 3AFC trial required understanding of the blue star to pick the correct word order among VSO, VOS, and SVO options. In the first generation, SVO was never a correct answer choice and participants were never taught an SVO sentence mapping. Participants received feedback in all but the final block of 3AFC trials. The scene-sentence mappings participants made here became the input for the next generation. Feedback consisted of a thumbs up upon answering a question correctly (positive) or a thumbs down upon answering a question incorrectly (negative), followed by the correct scene-sentence pair. No feedback participants saw their selected scene-sentence pair after every response to ensure equal language exposure. The six generations of a transmission chain all received the same type of feedback.

The Generalization part tested participants on novel stimuli mapped to the original language of 12 VSO and 12 VOS sentences. Feedback was not given.

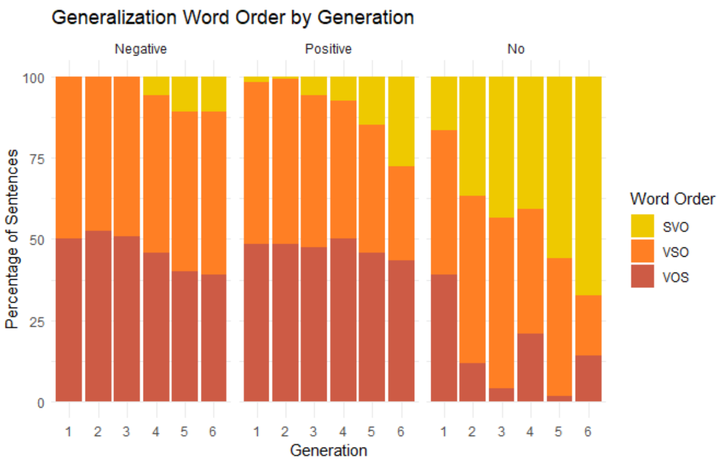


Fig 2. Participants maintained the original word orders best when feedback was given.

Results. Our preregistered sample of 90 native English speakers (60 female, 22 male, 8 non-binary; mean age = 19.8), all Cornell University undergraduates, completed the experiment for course credit. The convergence to SVO in the no feedback condition shown in Figure 2 is consistent with our predictions. Negative feedback best retained the original word order, subverting our expectations about positive feedback (cf. Frinsel et al., 2024). These findings provide new insight into the possible impact of generational transmission and feedback on stability of flexible word order in the cultural evolution of language.

References

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

Carroll, S., & Swain, M. (1993). Explicit and implicit negative feedback: An empirical study of the learning of linguistic generalizations. *Studies in Second Language Acquisition*, *15*(3), 357–386.

Ellis, R., Loewen, S., & Erlam, R. (2006). Implicit and explicit corrective feedback and the acquisition of L2 grammar. *Studies in Second Language Acquisition*, *28*(02), 339–368.

Frinsel, F.F., Trecca, F. & Christiansen, M.H. (2024). The role of feedback in the statistical learning of language-like regularities. *Cognitive Science, 48,* e13419.

Jeuniaux, P., Dale, R., & Louwerse, M. M. (2009). the role of feedback in learning form-meaning mappings. *In N. Taatgen, H. van Rijn, J. Nerbonne & L. Schomaker (Eds.), Proceedings of the 31st Annual Cognitive Science Society Conference* (pp. 1488–1493). *Austin, TX: Cognitive Science Society.*

Kirby, S., Cornish, H., & 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.

Levshina, N. (2021). Cross-linguistic trade-offs and causal relationships between cues to grammatical subject and object, and the problem of efficiency-related explanations. *Frontiers in Psychology*, *12*, 648200.

Monaghan, P., Ruiz, S., & Rebuschat, P. (2021). The role of feedback and instruction on the cross-situational learning of vocabulary and morphosyntax: Mixed effects models reveal local and global effects on acquisition. *Second Language Research*, *37*(2), 261–289.

Theisen-White, C., Kirby, S., & Oberlander, J. (2011). Integrating the horizontal and vertical cultural transmission of novel communication systems. *In L. Carlson, C. Hölscher, & T. Shipley (Eds.), Proceedings of the 33rd Annual Conference of the Cognitive Science Society (pp. 956-961). Austin, TX: Cognitive Science Society.*

The evolution of silence: the role of inter-turn speech pauses in the co-evolution of language and cooperation

Theresa Matzinger[*1,2], Michael Pleyer[3], Elizabeth Qing Zhang[4], and Przemysław Żywiczyński[3]

[*]Corresponding Author: theresa.matzinger@univie.ac.at

[1]Department of English and American Studies, University of Vienna, Vienna, Austria

[2]Cognitive Science Hub, University of Vienna, Vienna, Austria

[3]Center for Language Evolution Studies, Nicolaus Copernicus University in Toruń, Toruń, Poland

[4]School of Linguistic Sciences and Arts, Jiangsu Normal University, Xuzhou, China

Language and cooperation are closely intertwined aspects of human social interaction, with language serving as an important tool in the assessment of potential cooperation partners (Fox Tree, 2002; Brennan & Williams, 1995). Since pauses between turns of a conversation perform important social-communicative functions, we argue that pauses and their perception may have played a significant role in the co-evolution of language and cooperation.

We support this with data from an experiment testing how listeners use the duration of inter-turn speech pauses to assess others' knowledge, confidence and willingness to grant requests – all of which are highly relevant when evaluating others as potential cooperation partners (Authors, 2023; Authors, *in prep*). We hypothesized that, in general, speakers making long pauses would be regarded as less apt and willing (Roberts & Francis, 2013) but that listeners would be more tolerant towards long pauses in non-native speakers. This is because in non-native speakers, long pauses may result from prolonged cognitive processing when answering in a non-native language (Cenoz, 2000; Guyer et al., 2019) rather than from a lack of knowledge or willingness. Crucially, since evaluating others' cooperativeness is important across cultures, and pause production and perception are similar cross-linguistically (Matzinger & Fitch, 2021), we predicted similar effects across languages.

In our experiment, 100 native Polish-speaking raters listened to short staged conversations, during which a speaker asked questions or made requests that were answered or granted by either native speakers of Polish or native Chinese-speaking non-native speakers of Polish. The pauses before the answers were manipulated to be either short (200 ms) or long (1200 ms; cf. Roberts & Francis, 2013; Dingemanse & Liesenfeld, 2022). After listening, the raters judged each respondent on their knowledge, confidence and willingness. To test for cross-linguistic similarities, we replicated the experiment with Chinese raters and the two languages reversed.

Our results suggest similarities across languages and cultures: as predicted, Polish and Chinese raters perceived native speakers making long pauses as less knowledgeable, confident and willing (Fig. 1; see non-overlapping 95% confidence intervals). Also, for both rater groups, linear mixed effects models revealed a mediating effect of non-native accent on perceived willingness (interaction: *pause duration* * *accent*: $p < 0.001$ each), but not on knowledge and confidence. A potential reason for the difference between the findings on willingness versus knowledge and confidence is that requests may be more socially engaging and more directly relevant for interpersonal cooperative interactions than knowledge that reflects on partners' competence but not cooperativeness. The raters may therefore have evaluated willingness more carefully.

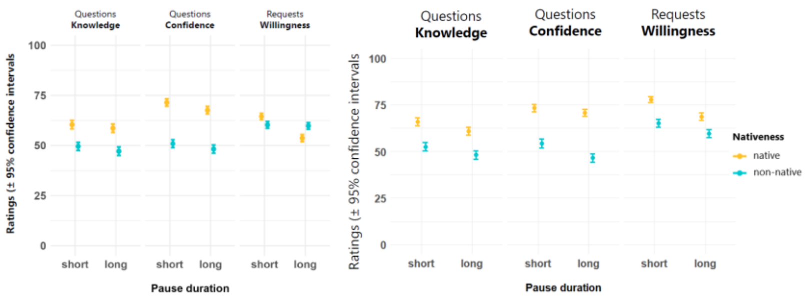


Figure 1. Mean ratings (± 95%-confidence intervals) of perceived knowledge, confidence and willingness when native vs. non-native speakers made short vs. long inter-turn pauses. Left: Polish raters, right: Chinese raters.

To further support the point that pauses may have been used as communicative signals in early language evolutionary stages when fully-developed shared linguistic communication systems have not yet been established, we will also show results from a follow-up study testing whether similar relationships hold if participants don't know the language spoken in the conversations. We predict that understanding the content of the answers should not play a pivotal role.

Finally, since pauses in turn-taking occur across species (Levinson, 2016), we will also discuss inter-turn pauses as promising candidates to investigate links between communication and cooperativeness across species.

Acknowledgements

Michael Pleyer was supported by project No. 2021/43/P/HS2/02729 co-funded by the National Science Centre and the European Union Framework Programme for Research and Innovation Horizon 2020 under the Marie Skłodowska-Curie grant agreement No. 945339.

References

Brennan, S. E. & Williams. M. (1995). The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language*, *34*, 383-398.

Cenoz, J. (2000). Pauses and hesitation phenomena in second language production. *ITL: Review of Applied Linguistics, 127*, 53-69.

Dingemanse, Mark & Andreas Liesenfeld. (2022). From text to talk: Harnessing conversational corpora for humane and diversity-aware language technology. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics 1 (Long Papers)* (pp.5614-5633). Dublin: Association for Computational Linguistics.

Fox Tree, J. E. (2002). Interpreting pauses and ums at turn exchanges. *Discourse Processes, 34,* 37–55.

Guyer, J. J., Fabrigar, L. R. & Vaughan-Johnston, T. I. (2019). Speech rate, intonation, and pitch: Investigating the bias and cue effects of vocal confidence on persuasion. *Personality and Social Psychology Bulletin, 45,* 389-405.

Levinson, S. C. (2016). Turn-taking in human communication–origins and implications for language processing. *Trends in Cognitive Sciences*, 20(1), 6-14.

Matzinger, T., & Fitch, W. T. (2021). Voice modulatory cues to structure across languages and species. *Philosophical Transactions of the Royal Society B, 376(1840)*, 20200393.

Roberts, F. & Francis, A. L. (2013). Identifying a temporal threshold of tolerance for silent gaps after requests. *The Journal of the Acoustical Society of America, 133*, EL471-EL477.

Rats (*Rattus norvegicus*) detect temporal rather than melodic changes

Ferran Mayayo[*1], Juan M. Toro[1,2]

[*]Corresponding Author: ferran.mayayo@upf.edu
[1] Center for Brain & Cognition, Universitat Pompeu Fabra, Barcelona, Spain
[2] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

Background. Music and language both have a rhythmic organization, and these fields even share some common rhythmic attributes. In music, meter is one of the fundamental components of rhythm. A metrical structure is perceived when some of the events are felt as more prominent than the others. Meter can be physically evoked by changes in the events (beats) composing a sequence, but it can also be induced by the listener over sequences of regular events. Even though humans readily perceive meter over rhythmic structures, it is not yet known to what extent meter perception is already present in non-human animals. In fact, the "vocal learning and rhythmic synchronization hypothesis" (Patel, 2006) suggests that only vocal learning species (animals that are able to acquire non-innate sounds; Jarvis, 2004) might have the required processing abilities as to induce meter in rhythmic sequences.

Aims. Our aim is to explore the extent to which the basis for meter perception can be found in non-human animals that are not vocal learners and cannot modify their vocal output, such as the rat. We thus want to test whether rats detect and discriminate sequences with different metrical structures shaped by melodic and temporal accents.

Method. 40 naive rats were familiarized with auditory rhythmic sequences that evoked a specific metrical structure. They were individually placed in isolated response boxes with a speaker, a nose-poking detector and a pellet feeder. The animals were trained to poke their nose into the feeder to receive food reward. Once they learned this association, we ran 30 familiarization sessions. During

familiarization, the rats were rewarded with food when poking after stimuli presentation. In test sessions, the animals were presented with 24 familiarization sequences and 16 test sequences (8 familiar sequences and 8 novel sequences). Total number of responses (nose-poking) to familiar and to novel test sequences was analyzed. Importantly, neither familiar test nor novel test sequences were rewarded. Each experiment lasted 2 months approximately. The animals were caged in pairs in a quiet environment.

In a first experiment, we familiarized rats with rhythmic sequences with a duple or triple metrical structure, which was defined by melodic accents. During test sessions they had to discriminate between familiar and novel sequences with different metrical structures (duple vs. triple meter), between familiar and novel isotonic sequences (all tones set at the same pitch frequency) and between familiar and novel non-isochronous sequences (different duration between interval onsets). In a second experiment, rats were familiarized with rhythmic sequences with a long-short tone pattern (LS) or a long-short-short tone pattern (LSS). Each pattern induced a different metrical structure, shaped by temporal accents. In tests, rats were presented with novel temporal sequences. They had to discriminate between rhythmic sequences with different metrical structures, but also between sequences with different rhythmic groupings and between sequences with different time interval ratios.

Results. In Experiment 1, we observed that the animals did not discriminate between duple and triple meter sequences, nor did they discriminate between duple or triple meter sequences and isotonic sequences. Conversely, they were able to discriminate between isochronous, metrical sequences and non-isochronous sequences. Thus, rats probably focused on the temporal structure of the stimuli while disregarding melodic changes. In Experiment 2, the results showed that rats can discriminate between LS and LSS sequences (with constant time interval ratio), between familiar and novel LS sequences or familiar and novel LSS sequences (with constant grouping), and between LSS and LSSSS sequences (with constant meter).

Conclusions. Together, the results from the 2 experiments suggest that rats, a non-vocal learning mammal, can detect differences between rhythmic sequences if changes are present on a temporal level rather than a melodic or tonal level. We did not observe concluding evidence regarding the rats' ability to process meter, but the results suggest that metrical structures with temporal accents (rather than dynamic or melodic accents) would be more informative for the perception of metrical structures in rats.

References

Jarvis, E. D. (2004). Learned birdsong and the neurobiology of human language. *Annals of the New York Academy of Sciences, 1016*, 749–777.

Patel, A. (2006). Musical rhythm, linguistic rhythm, and human evolution. *Music Perception, 24*, 99–104.

A complex systems perspective on language evolution

Jérôme Michaud*[1,2]

*Corresponding Author: jerome.michaud@mdu.se
[1]Centre for Cultural Evolution, Department of Psychology, Stockholm University, Stockholm, Sweden
[2]Department of Mathematics and Physics, Mälardalen University, Västerås, Sweden

Language is acknowledged as a complex adaptive system, but its implications from a modeling perspective remain largely under-explored. This paper explores the application of complex systems theory, as formalized by Thurner, Hanel, and Klimek in their book, "Introduction to the Theory of Complex Systems" (2018), to the study of language evolution. Here we show that a simple agent-based model of language evolution that incorporates innovations, combinations, and the social transmission of cultural traits is sufficient to produce a rapid complexification of the linguistic systems in both evolutionary and developmental contexts. A key feature of this model, necessary for the emergence of complexity, is the ability to productively combine and selectively filter out traits. This highlights the importance of combinatorial interactions in the self-organization of linguistic systems. From a developmental perspective, it stresses the importance of acquiring a sufficient number of traits to bootstrap the complexity of natural language.

1. Introduction

As many scholars have observed, language is a complex adaptive system (Beckner et al., 2009; Ellis & Larsen-Freeman, 2009; Ellis, 2011; Schmid, 2020). Despite this recognition, the formalization of this complex nature remains an under-explored territory in the field of language evolution. One prevalent aspect of complex adaptive systems is that they self-organize. When it comes to language, self-organization occurs both at the individual level and at the population level (Schmid, 2020; de Boer, 2011). The Conventionalization-Entrenchment (CE) model of Schmid (Schmid, 2020) clearly points to the separation of the social dynamics, i.e. conventionalization, and the individual dynamics, i.e. entrenchment. In (de Boer, 2011), self-organizing processes are also categorized as individual or population processes. When it comes to modelling these self-organizing processes, most work have focused on self-organization at the population level using agent-based models (Hurford, 1987; Steels, 1995, 1997; Batali, 1998; Kirby, 2002). In contrast, self-organizing processes at the individual level have been given less attention and mostly focus on the self-organization of sound systems such as vowel systems (Lijencrants & Lindblom, 1972; Schwartz, Boë, Vallée,

& Abry, 1997; Ke, Ogura, & Wang, 2003) or syllable systems (Lindblom, MacNeilage, & Studdert-Kennedy, 1984). There is a need for modelling the self-organization of grammatical linguistic traits into systems to account for observed linguistic trait dependencies (Greenberg et al., 1963; Croft, 2002; Dunn, Greenhill, Levinson, & Gray, 2011; Skirgård et al., 2023; Enfield, 2017).

In this paper, we will use the theory of complex systems of Thurner et al. (2018), which relies on studying the joint dynamics of a number of entities and their interactions, to model the self-organization of linguistic traits into coherent linguistic systems. Models of cultural systems have recently been developed (Friedkin, Proskurnikov, Tempo, & Parsegov, 2016; Buskell, Enquist, & Jansson, 2019; Yeh, Fogarty, & Kandler, 2019; Goldberg & Stein, 2018), but have often fallen short in capturing the combinatorial nature of trait interactions. In contrast, the field of biology has explored combinatorial models of evolution extensively (Jain & Krishna, 2001; Solé & Manrubia, 1996; Kauffman, 1993). Notably, the open-ended co-evolving combinatorial critical model (CCC model) (Klimek, Thurner, & Hanel, 2010; Thurner et al., 2018) stands out in providing a simple model of evolution capturing many key properties of complex evolutionary systems and stressing the importance of both productive and destructive combinatorial interactions.

The main idea of this paper is to use the CCC model as an idealized model of a linguistic system in which traits represents linguistic features, such as word order or the presence of prepositions, and embed it into socially interacting individuals, thus providing a multi-agent version of the CCC model and capturing self-organizing processes both at the individual and at the population level. Our proposed model offers insights into the emergence and subsequent evolution of complex languages and displays punctuated equilibria in both evolutionary and developmental contexts. The non-linear dynamics of this model is highlighted by a phase transition from low to high diversity, potentially shedding light on the first emergence of complex languages. Additionally, we observe that the evolution of individual agents exhibits a pattern resembling human linguistic development, with a learning period followed by a phase of high diversity.

2. Complex systems

According to Thurner et al. (2018), complex systems can be conceptualized as co-evolving multilayer networks whose nodes represent various types of entities labeled by Latin indices i and links represents various types of interactions labeled by Greek indices α, with each interaction type defining a distinct layer of the network. Importantly, interactions are not static but dynamically evolve over time.

The strength of interaction α between elements $i, j, k, \ldots$ at time t is denoted as $M^{\alpha}_{ijk\ldots}(t)$. Interactions often involve more than two entities, encoding combinatorial interactions. Elements themselves are characterized by states denoted as $\sigma_i(t)$. In complex systems, states and interactions are not independent; rather, they mutually influence each other, resulting in a phenomenon known as *co-evolution*. This notion of co-evolution should not be confused with that of gene-culture co-evolution or other similar interpretation. Due to the discrete nature of interactions between interaction networks and states, complex systems are inherently *algorithmic*, making them challenging to describe analytically. Formally, co-evolving multiplex networks can be expressed as:

$$\begin{aligned} \sigma_i(t+dt) &\sim \sigma_i(t) + F(M^{\alpha}_{ijk\ldots}(t), \sigma_j(t)) \\ M^{\alpha}_{ijk\ldots}(t+dt) &\sim M^{\alpha}_{ijk\ldots}(t) + G(M^{\alpha}_{ijk\ldots}(t), \sigma_j(t)) \end{aligned}, \tag{1}$$

where F and G are algorithms governing the computation of the next iteration.

The complexity that can arise from evolution is unlimited, but what is accessible from a given state of the system is usually much smaller. To account for that, Stuart Kauffman introduced the concept of the *adjacent possible* (Kauffman, 1969, 1993). The adjacent possible represents the set of all potential states of the world that could potentially exist in the subsequent time step and encapsulates the path-dependency inherent in evolutionary processes. In addition, most complex systems are both robust and adaptive, two properties that rarely occur together. In any dynamical system, the boundary separating the region of stability from that of chaotic behaviour in the phase space is termed the *edge of chaos* (Lewin, 1999). Evolutionary systems are often *self-organized critical systems* (Bak, Tang, & Wiesenfeld, 1987) meaning that their dynamics drives them toward the edge of chaos. At the edge of chaos, a system is metastable and alternates between stable and chaotic phases, leading to punctuated equilibria.

3. Modelling language as a complex evolutionary system

In language evolution, we study how large groups of speakers and their linguistic traits change over time. Therefore, we identify two principal categories of entities: (i) linguistic traits that collectively compose a language, and (ii) the speakers of the language, representing the individual and population levels respectively (de Boer, 2011; Schmid, 2020). We will use the CCC model (Klimek et al., 2010; Thurner et al., 2018) as a model of individual processes and use an agent-based approach to model transmission and conventionalization.

3.1. *The CCC model and self-organization at the individual level*

The CCC model captures complex evolutionary systems where entities, denoted by i, have binary states represented as $\sigma_i(t)$, where 1 indicates presence at time t, and 0 indicates absence. Entities interact through two types of interactions: constructive (M^{+}_{ijk}) and destructive (M^{-}_{ijk}) combinatorial interactions. When $M^{+}_{ijk} = 1$, it signifies that the joint presence of entities i and j supports the emergence of entity k, while $M^{-}_{ijk} = 1$ implies that their joint presence inhibits entity

k. For simplicity, we assume that an entity engages in r^+ constructive interactions and r^- destructive interactions on average. The fitness ($f_k(t)$) of entity k at time t is calculated as:

$$f_k(t) = \sum_{i,j} \left(M^+_{ijk} - M^-_{ijk} \right) \sigma_i(t)\sigma_j(t), \tag{2}$$

measuring the difference between constructive and destructive interactions. If $f_k(t) > 0$, entity k is present in the next time step; if $f_k(t) < 0$, it's absent. When $f_k(t) = 0$, the entity's state remains unchanged. Additionally, with a probability p, the state of entity k is randomly flipped, representing innovation. The CCC model exhibits common characteristics of evolutionary complex systems, including punctuated equilibria and a low-to-high diversity phase transition.

Here, entities represent linguistic features such as word order of the presence of prepositions. The elements $M^{\pm}_{ijk}$ represent their interactions and can be conceived as idealized version of Greenberg's universals (Greenberg et al., 1963). For example, if we consider the typological universal: *If a language has dominant SOV order and the genitive follows the governing noun, then the adjective likewise follows the noun* we could model such a situation by considering trait i to be a *SOV word order*, trait j to be that *the genetive follows the governing noun*, and trait k to be that *the adjective follows the noun*, then M^+_{ijk} encodes the fact that together traits i and j support trait k.

3.2. *Social learning and self-organization at the population level*

In order to model social learning, we assume that each agent is an instance of the CCC model. Trait ownership is now encoded as an interaction between agents and traits and denoted by $\sigma^{(i)}_k(t)$, where 1 indicates that agent i possesses trait k, and 0 indicates the absence of the trait. Interaction tensors between traits are defined as in the CCC model and we can define the fitness of an agent i as

$$f^{(i)}_k(t) = \sum_{l,m} \left(M^+_{lmk} - M^-_{lmk} \right) \sigma^{(i)}_l(t)\sigma^{(i)}_m(t)\,. \tag{3}$$

This fitness is different for every agents and is an example of a *subjective fitness*. The concept of subjective fitness in language dynamics is discussed in (Michaud, 2019, 2020). This models takes into account both evolution and self-organization of the linguistic system.

In addition, the age of an agent at time t is encoded as $a^{(i)}(t)$. The model is driven by social learning, i.e. copying a trait from another random agent (we assume homogeneous population for simplicity), which can formally be written as $\sigma^{(i)}_k(t+1) = \sigma^{(j)}_k(t)$. When acquiring a new trait, an agent performs n_C steps of the CCC model to stabilize their internal system. In addition, with probability

p, their will innovate a new trait. This model has two different timescales, that of social interactions and that of internal self-organization.

A simple population dynamics is added. Agents are initialized without any traits and gradually acquire them as they age. The population resides in an environment with a fixed carrying capacity. Agents face a constant probability of death, with new individuals replacing those who perish. Furthermore, agents cannot age beyond a predefined maximum age. Upon an agent's death, the corresponding elements of $\sigma^{(i)}(t)$ are reset to 0, modelling the birth of a new agent. Destructive interactions models filtering of traits not fitting within the system.

4. Results

We conducted simulations starting with a population devoid of any linguistic traits to model both language emergence and language evolution. Traits emerge through the innovation rule of the CCC dynamics. Once a trait has emerged, it can be adopted by other agents and combined with other traits, either supporting or inhibiting the emergence of additional traits.

Figure 1 depicts a typical simulation run with a population of approximately 100 agents, an expected lifespan of 50 time steps, and a maximum age of 100. For the CCC step, we considered 100 possible traits, with $r^+ = 3$, $r^- = 4$, and $n_C = 10$, and an innovation probability of $p = 3 \cdot 10^{-6}$. These parameters result in an innovation event, on average, every 3 time steps. In the top left panel, we observe the population's evolution, beginning with low trait diversity. After two plateaus, the average number of traits quickly increases and stabilizes at around 30. The top right panel illustrates ontogeny at the end of the simulation, showing a sharp rise in the number of traits followed by stabilization. To provide deeper insights into this process, the bottom left panel displays the learning history of a representative agent. It shows a period of low trait diversity transitioning to high diversity when a sufficient number of traits have been acquired. The curve in the top right panel doesn't exhibit a jump because the shift from low to high diversity occurs at different ages for different agents. Once agents enter the high diversity phase, we observe punctuated equilibria, a characteristic of the CCC model. The final panel presents the increase in diversity with r^+. It shows that below $r_+ = 1.5$ no diversity emerges in the population and then it is roughly proportional to r^+.

5. Discussion

One of the big question in language evolution is how complex language first appeared, and why only humans have it. In this paper, we suggest a vital part of the answer is our ability to combine linguistic traits and to organize them into systems. This combinatorial aspect of language is pervasive, extending from phonology (Lijencrants & Lindblom, 1972; Lindblom et al., 1984), to morpho-syntax (Greenberg et al., 1963; Croft, 2002; Enfield, 2017), to semantics (Wolfe, 1972;

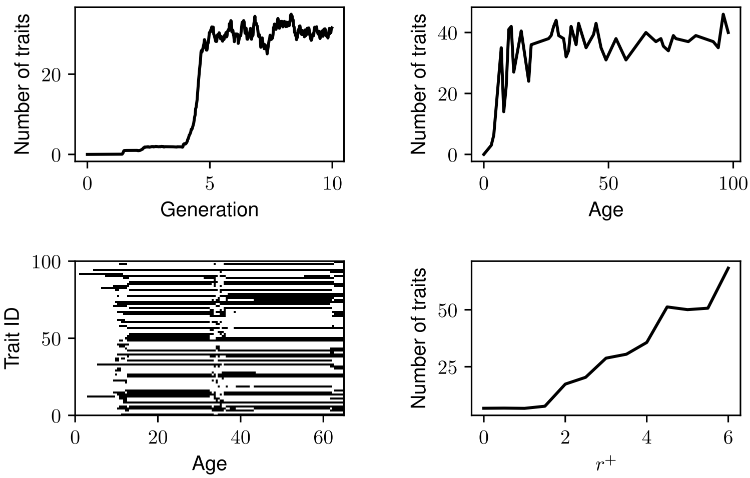


Figure 1. Top left panel illustrates the jump in diversity (average number of traits in the population) after about 5 generations. Top right panel shows the ontogeny of the number of traits at the end of the simulation. Bottom left panel shows the learning history of a typical agent (black indicates trait presence). Bottom right panel shows the averaged diversity achieved over 10 simulation runs when varying r^+ and keeping $r^- = 4$ constant.

de Boer, 2005; Wierzbicka, 2015). We argue that, for the emergence of complex language, speakers must possess the ability to productively combine linguistic traits as well as the ability to filter out traits not fitting into the system. We use Greenberg's implicational universals as an illustration of such processes. This leads to the hypothesis that humans may be better at productively combining traits than other animals and is compatible with the sequence hypothesis that states that only humans perceive order faithfully (Jon-And, Jonsson, Lind, Ghirlanda, & Enquist, 2023; Enquist, Ghirlanda, & Lind, 2023; Lind, Vinken, Jonsson, Ghirlanda, & Enquist, 2023; Ghirlanda, Lind, & Enquist, 2017).

From a developmental perspective, our model stresses the need for learners to acquire a critical mass of traits to bootstrap the complexity observed in later stages of life. This observation hints at the presence of ”keystone traits” (Thurner et al., 2018) that initiate cascades toward complexity, marking the transition from childhood (a low diversity phase) to adulthood (a high diversity metastable phase) in linguistic development.

In conclusion, our exploration of language evolution using the multi-agent CCC model offers valuable insights into the joint dynamic between linguistic traits and language users. By simulating the emergence, evolution and self-organization of traits within a population, we provide a framework for understanding the complex, punctuated, and socially influenced nature of language evolution.

Acknowledgements

This work was supported by the Marianne and Marcus Wallenberg Foundation (grant no. 2021.0039) and by the Swedish Research Council (VR 2022-02737).

References

Bak, P., Tang, C., & Wiesenfeld, K. (1987). Self-organized criticality: An explanation of the 1/f noise. *Physical review letters*, *59*(4), 381.

Batali, J. (1998). Computational simulations of the emergence of grammar. *Approaches to the Evolution of Language-Social and Cognitive Bases-*.

Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., Holland, J., Ke, J., Larsen-Freeman, D., et al.. (2009). Language is a complex adaptive system: Position paper. *Language learning*, *59*, 1–26.

Buskell, A., Enquist, M., & Jansson, F. (2019). A systems approach to cultural evolution. *Palgrave Communications*, *5*(1), 1–15.

Croft, W. (2002). *Typology and universals.* Cambridge University Press.

de Boer, B. (2005). Self-organisation in language. In C. Hemelrijk (Ed.), *Self-organisation and evolution of biological and social systems* (p. 123–139). Cambridge University Press.

de Boer, B. (2011). Self-organization and language evolution. In *The Oxford Handbook of Language Evolution.* Oxford University Press.

Dunn, M., Greenhill, S. J., Levinson, S. C., & Gray, R. D. (2011). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, *473*(7345), 79–82.

Ellis, N. C. (2011). The emergence of language as a complex adaptive system. *The Routledge handbook of applied linguistics*, *666679*.

Ellis, N. C., & Larsen-Freeman, D. (2009). *Language as a complex adaptive system* (Vol. 11). John Wiley & Sons.

Enfield, N. (2017). *Dependencies in language: On the causal ontology of linguistic systems.* Language Science Press.

Enquist, M., Ghirlanda, S., & Lind, J. (2023). *The human evolutionary transition: From animal intelligence to culture.* Princeton University Press.

Friedkin, N. E., Proskurnikov, A. V., Tempo, R., & Parsegov, S. E. (2016). Network science on belief system dynamics under logic constraints. *Science*, *354*(6310), 321–326.

Ghirlanda, S., Lind, J., & Enquist, M. (2017). Memory for stimulus sequences: a divide between humans and other animals? *Royal Society open science*, *4*(6), 161011.

Goldberg, A., & Stein, S. K. (2018). Beyond social contagion: Associative diffusion and the emergence of cultural variation. *American Sociological Review*, *83*(5), 897–932.

Greenberg, J. H., et al.. (1963). Some universals of grammar with particular reference to the order of meaningful elements. *Universals of language*, *2*, 73–113.

Hurford, J. (1987). *Language and number: the emergence of a cognitive system.* Oxford: Blackwell.

Jain, S., & Krishna, S. (2001). A model for the emergence of cooperation, interdependence, and structure in evolving networks. *Proceedings of the National Academy of Sciences*, *98*(2), 543–547.

Jon-And, A., Jonsson, M., Lind, J., Ghirlanda, S., & Enquist, M. (2023). Sequence representation as an early step in the evolution of language. *PLOS Computational Biology*, *19*(12), e1011702.

Kauffman, S. A. (1969). Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of theoretical biology*, *22*(3), 437–467.

Kauffman, S. A. (1993). *The origins of order: Self-organization and selection in evolution.* Oxford University Press, USA.

Ke, J., Ogura, M., & Wang, W. S.-Y. (2003). Optimization models of sound systems using genetic algorithms. *Computational Linguistics*, *29*(1), 1–18.

Kirby, S. (2002). Natural language from artificial life. *Artificial life*, *8*(2), 185–215.

Klimek, P., Thurner, S., & Hanel, R. (2010). Evolutionary dynamics from a variational principle. *Phys Rev E*, *82*.

Lewin, R. (1999). *Complexity: Life at the edge of chaos.* University of Chicago Press.

Lijencrants, J., & Lindblom, B. (1972). Numerical simulations of vowel quality systems: The role of perceptual contrasts. *Language*, *48*, 839–862.

Lind, J., Vinken, V., Jonsson, M., Ghirlanda, S., & Enquist, M. (2023). A test of memory for stimulus sequences in great apes. *Plos one*, *18*(9), e0290546.

Lindblom, B., MacNeilage, P., & Studdert-Kennedy, M. (1984). Self-organizing processes and the explanation of phonological universals. 181-203.

Michaud, J. (2019). Dynamic preferences and self-actuation of changes in language dynamics. *Language Dynamics and Change*, *9*(1), 61–103.

Michaud, J. (2020). A game theoretic perspective on the utterance selection model for language change. In A. Ravignani, C. Barbieri, M. Martins, M. Flaherty, Y. Jadoul, E. Lattenkamp, H. Little, K. Mudd, & T. Verhoef (Eds.), *The evolution of language: Proceedings of the 13th international conference (evolang13).*

Schmid, H.-J. (2020). *The dynamics of the linguistic system: Usage, conventionalization, and entrenchment.* Oxford University Press.

Schwartz, J.-L., Boë, L.-J., Vallée, N., & Abry, C. (1997). The dispersion-focalization theory of vowel systems. *Journal of phonetics*, *25*(3), 255–286.

Skirgård, H., Haynie, H. J., Blasi, D. E., Hammarström, H., Collins, J., Latarche, J. J., Lesage, J., Weber, T., Witzlack-Makarevich, A., Passmore, S., et al.. (2023). Grambank reveals the importance of genealogical constraints on

linguistic diversity and highlights the impact of language loss. *Science Advances*, *9*(16), eadg6175.
Solé, R. V., & Manrubia, S. C. (1996). Extinction and self-organized criticality in a model of large-scale evolution. *Physical Review E*, *54*(1), R42.
Steels, L. (1995). A self-organizing spatial vocabulary. *Artificial life*, *2*(3), 319–332.
Steels, L. (1997). Self-organizing vocabularies. *Artificial Life*, *5*, 179–184.
Thurner, S., Hanel, R., & Klimek, P. (2018). *Introduction to the theory of complex systems.* Oxford University Press.
Wierzbicka, A. (2015). Natural semantic metalanguage. *The international encyclopedia of language and social interaction*, 1–17.
Wolfe, P. M. (1972). *Linguistic change and the great vowel shift in english.* Univ of California Press.
Yeh, D. J., Fogarty, L., & Kandler, A. (2019). Cultural linkage: the influence of package transmission on cultural dynamics. *Proceedings of the Royal Society B*, *286*(1916), 20191951.

Visual communication in heterogeneous populations of artificial communicating agents

Daniela Mihai*[1]

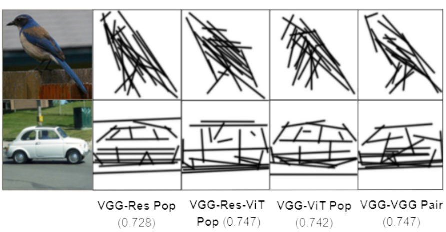
*Corresponding Author: adm1c22@soton.ac.uk
[1]Vision, Learning and Control Group, University of Southampton, Southampton. UK

Innovations in artificial neural networks, deep and reinforcement learning techniques have led to research on communication protocols emergent from multi-agent interactions in the form of gameplay (Chaabouni et al., 2020; Guo, 2019; Lazaridou et al., 2018; Havrylov & Titov, 2017; Nölle et al., 2018; De Boer & Zuidema, 2010). Referential communication games (Lewis, 1969) have often been used as a simple, yet effective task that allows agents to develop their language through cooperation. These communication protocols evolved by artificial agents can be of various types, discrete or continuous, symbolic as in token-based (Havrylov & Titov, 2017) or iconic (Mihai & Hare, 2021). Recent studies looked into the effect of population size (Chaabouni et al., 2022) and heterogeneity (Rita et al., 2022; Mahaut et al., 2023) on token-based communication protocols.

In this work, we explore how population heterogeneity impacts a visual communication protocol of agents interacting through drawing. Using the drawing game environment of Mihai and Hare (2021), we study populations of agents with different pretrained visual encoding networks. Different model architectures are used as a proxy for agent heterogeneity to account for different visual experiences during the lifetime (see the supplementary for further discussion). The communication is through sketches made of 20 black lines. We explore how the graphical protocol changes under social pressures as agents are part of a community and need to adapt to more communicating partners. Previous works showed that pairs of agents playing a referential game tend to develop symbolic representations of the world specific to their interaction partner (Hawkins et al., 2023). Fay et al. (2014) also suggest that iconic representations become symbolic through repetitive interaction. Conversely, if the graphical communication evolves as the population stochastically participates in the games, then all participants will shape the final protocol. We hypothesise that the drawings of agents from population-based training will be more generalised across the population, and hence more iconic and less abstract than those emerging from homogeneous pair-wise interactions.

The setup involves a sender communicating through drawing about an image from STL-10 dataset (Coates et al., 2011). The game's goal is, based on the sketch that the sender produces, for the receiver to correctly distinguish the target image from 50 candidates. The populations we test are of 4 and 6 agents (i.e. 2, or 3 respectively, senders and receivers) to sample from at each time step. It is worth noting that what we refer to as a sender agent can only produce drawings, and a receiver only interprets. Depending on the population, an agent can be instantiated with one of the 3 visual feature extraction modules: VGG16 (Simonyan & Zisserman, 2015), ResNet18 (He et al., 2016) and Vision Transformer (ViT) (Radford et al., 2021). The features are extracted after the last convolution layer and are passed through an additional batch normalisation layer before being fed into the sketch drawing module. The population model architecture, detailing each agent's learnable and pretrained modules can be found in the supplementary.

Preliminary experiments show that some populations are more successful than others. For example, the population of agents with VGG and ViT visual systems overall better solve the task. Due to the stochastic nature of the training, the communication rate can vary considerably throughout training as different agents interact and establish the convention at each step. In Fig. 1 Right we report average population accuracy, i.e. the task success averaged across all possible agent pairs in the population tested on the same evaluation set. In the supplementary material, we compare the communication success of heterogeneous populations with that of homogeneous pairs and observe the former are more difficult to train. Although additional training steps can sometimes improve average task success over the population, the iconicity of sketches does not significantly change.



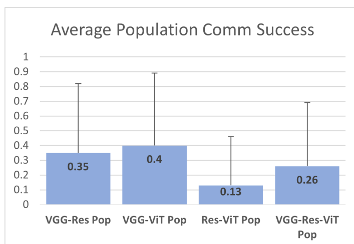


Figure 1. Left: Sketches produced by a VGG-sender agent trained in different population/pair configurations with *perceptual similarity* score as a measure of iconicity - lower means more iconic. Right: Average communication success of different populations on the test set after 250 epoch-training. The line represents the standard deviation from the mean accuracy across pairs in the population.

Fig. 1 Left compares sketches and reports perceptual similarity, as defined by Zhang et al. (2018), between the target image and sketches produced by VGG-senders trained in different configurations. This measure computed across deep features of a pretrained network is used as a proxy for iconicity - the more similar the representation (sketch) is to the real object (image), the more iconic (Peirce, 1867). These preliminary results point towards the hypothesis that sketches produced by population agents are more *iconic* than those of homogeneous agents.

References

Chaabouni, R., Kharitonov, E., Bouchacourt, D., Dupoux, E., & Baroni, M. (2020). Compositionality and generalization in emergent languages. *arXiv preprint arXiv:2004.09124*.

Chaabouni, R., Strub, F., Altché, F., Tarassov, E., Tallec, C., Davoodi, E., Mathewson, K. W., Tieleman, O., Lazaridou, A., & Piot, B. (2022). Emergent communication at scale. In *International conference on learning representations.*

Coates, A., Ng, A., & Lee, H. (2011). An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics* (pp. 215–223).

De Boer, B., & Zuidema, W. (2010). Multi-agent simulations of the evolution of combinatorial phonology. *Adaptive Behavior*, *18*(2), 141–154.

Fay, N., Ellison, M., & Garrod, S. (2014). Iconicity: From sign to system in human communication and language. *Pragmatics & Cognition*, *22*(2), 244–263.

Guo, S. (2019). Emergence of numeric concepts in multi-agent autonomous communication. *arXiv preprint arXiv:1911.01098*.

Havrylov, S., & Titov, I. (2017). Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30* (pp. 2149–2159). Curran Associates, Inc.

Hawkins, R. D., Sano, M., Goodman, N. D., & Fan, J. E. (2023). Visual resemblance and interaction history jointly constrain pictorial meaning. *Nature Communications*, *14*(1), 1–13.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Lazaridou, A., Hermann, K. M., Tuyls, K., & Clark, S. (2018). Emergence of linguistic communication from referential games with symbolic and pixel input. In *International conference on learning representations.*

Lewis, D. K. (1969). *Convention: A philosophical study.* Wiley-Blackwell.

Mahaut, M., Franzon, F., Dessì, R., & Baroni, M. (2023). Referential communication in heterogeneous communities of pre-trained visual deep networks. *arXiv preprint arXiv:2302.08913*.

Mihai, D., & Hare, J. (2021). Learning to draw: Emergent communication through sketching. In *Advances in neural information processing systems.*

Nölle, J., Staib, M., Fusaroli, R., & Tylén, K. (2018). Environmental and social factors motivate the emergence of systematic categories and signs. In C. Cuskley, M. Flaherty, H. Little, L. McCrohon, A. Ravignani, & T. Verhoef (Eds.), *The evolution of language: Proceedings of the 12th international conference (evolangxii).* NCU Press.

Peirce, C. S. (1867). Five hundred and eighty-second meeting. may 14, 1867. monthly meeting; on a new list of categories. *Proceedings of the American Academy of Arts and Sciences*, *7*, 287–298.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).

Rita, M., Strub, F., Grill, J.-B., Pietquin, O., & Dupoux, E. (2022). On the role of population heterogeneity in emergent communication. In *The international conference on learning representations (iclr) 2022.*

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International conference on learning representations.*

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. *CoRR*, *abs/1801.03924*.

Iconicity and compositionality in emerging vocal communication systems: a Virtual Reality approach

Kotryna Motiekaitytė[*1], Oxana Grosseck [1,2], Luisa Wolf[1], Hans Rutger Bosker[2], David Peeters[3,4], Marcus Perlman[5], Gerardo Ortega[5], and Limor Raviv[1,6]

[*]Corresponding Author: kotryna.motiekaityte@mpi.nl
[1]LEADS group, Max Planck Institute for Psycholinguistics, Nijmegen, NL
[2]Donders Institute, Radboud University, Nijmegen, NL
[3]Dept. of Communication and Cognition, Tilburg University, Tilburg, NL
[4]Multimodal language dept., Max Planck Institute for Psycholinguistics, Nijmegen, NL
[5]Dept. of English Language and Linguistics, University of Birmingham, Birmingham, UK
[6]cSCAN, University of Glasgow, Glasgow, UK

A central discussion within the field of language evolution is whether human communication at its origins was spoken, gestured, or multimodal as we see today (Perniss, 2018). Since motivated signals can bootstrap language emergence and learning (Perniss & Vigliocco, 2014), the great affordance for iconicity in gesture and sign languages was seen as supportive of a gesture-first origin (Tomasello, 2010; Fay et al., 2014). However, recent typological and experimental studies suggest that the vocal modality can afford more iconicity than previously thought (Perlman, 2017). Other work suggests that while some meanings can be easily expressed iconically in both modalities (e.g., size, shape, speed), some meanings may be more readily expressed iconically in the gestural modality (e.g., spatial relations) or in the vocal modality, (e.g., qualities of sound) (for overview: Dingemanse et al., 2015; Perlman & Cain, 2014). At the same time, highly iconic communication systems may reduce the pressure to develop compositional structure or systematic form-to-meaning mapping (Verhoef et al., 2016).

Here we examine the development of iconicity and compositionality in novel communication systems across different modalities (vocal, gestural, and multimodal) and across different semantic features. We conducted a dyadic communication game in an immersive Virtual Reality environment, which allows for ecological validity while maintaining high experimental control (Nölle et al., 2020; Peeters, 2019). Namely, participants interacted face-to-face and without computer interference to refer to novel stimuli around them in a virtual forest, while not being allowed to use any existing language. We compared the creation of new communication systems in 18 pairs across three experimental conditions (vocalization-only, gesture-only, or multimodal; 6 pairs per condition). The stimuli consisted of 32 fantasy creatures that varied by four semantic features: shape (4 types of creatures, Figure 1), size (small vs. big), movement (walk vs. jump), and speed (fast vs. slow). Our results show that communicative success was always highest in the gesture condition, with over 90% accuracy throughout (replicating Fay et al., 2014; Macuch Silva et al., 2020). For vocal and multimodal conditions, accuracy started lower yet increased over time, reaching 73% and 80%, respectively, by the end.

Since annotations of the gesture condition are still ongoing, here we focus on the vocal modality. Specifically, we analyzed the degree of iconicity (measured as effects between semantic and acoustic features) and compositionality (measured as the pair-wise correlation between meanings and orthographic annotations) in the vocal condition. Each vocalization was coded for duration, pitch, loudness, harmonics-to-noise ratio, number of syllables, and speech rate using PRAAT, and data was analyzed using mixed effect regression models in R. Concerning iconicity, we found that multiple acoustic features were significantly related to semantic features. For example, compared to small creatures, big creatures were typically described using lower pitch or louder vocalizations – in line with work on sound symbolism (Nygaard et al., 2009; Perlman & Cain, 2014). Turning to compositionality, novel vocal systems showed varying degrees of compositional structure (Figure 1), with some pairs developing highly structured vocalizations. Compositional structure increased significantly over time and led to increasing communicative success. Finally, pairs that developed the most compositional systems did not necessarily rely less on iconicity for successful communication and vice versa, i.e., the least accurate pairs struggled to employ either iconicity or compositionality. Together, our results suggest that both iconicity and compositionality can develop and co-exist in newly emerging vocal communication systems.

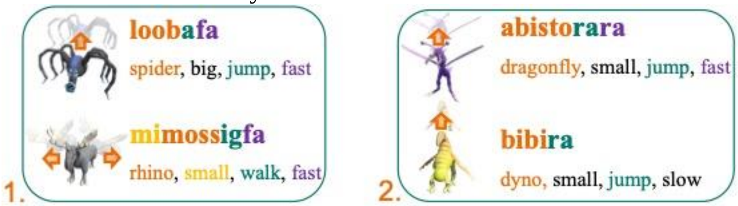


Figure 1. Examples of compositional structure in the annotated vocalizations of two different pairs.

References

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, iconicity, and systematicity in language. *Trends in cognitive sciences*, 19(10), 603-615.

Fay, N., Lister, C. J., Ellison, T. M., & Goldin-Meadow, S. (2014). Creating a communication system from scratch: gesture beats vocalization hands down. *Frontiers in Psychology*, 5, 354.

Macuch Silva, V., Holler, J., Ozyurek, A., & 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.

Nölle, J., Hartmann, S., & Tinits, P. (2020). Language evolution research in the year 2020: A survey of new directions. *Language Dynamics and Change,* 10(1), 3-26.

Nygaard, L. C., Herold, D. S., & Namy, L. L. (2009). The semantics of prosody: Acoustic and perceptual evidence of prosodic correlates to word meaning. *Cognitive science*, 33(1), 127-146.

Peeters, D. (2019). Virtual reality: A game-changing method for the language sciences. *Psychonomic bulletin & review*, 26, 894-900.

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.

Perlman, M., & Cain, A. A. (2014). Iconicity in vocalization, comparisons with gesture, and implications for theories on the evolution of language. *Gesture*, 14(3), 320-350.

Perniss, P. (2018). Why we should study multimodal language. *Frontiers in psychology*, 9, 1109.

Perniss, P., & 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.

reading. *Cognitive Science*, 39(6), 1348-1368.

Tomasello, M. (2010). *Origins of human communication*. MIT press.

Verhoef, T., Kirby, S., & De Boer, B. (2016). Iconicity and the emergence of combinatorial structure in language. *Cognitive science*, 40(8), 1969-1994.

Shared context helps maintain lexical variation

Katie Mudd*[1] and Marieke Schouwstra[1]

*katie.mudd@ai.vub.ac.be
[1]Institute for Logic, Language and Computation, University of Amsterdam, the Netherlands

How does the amount of social and psychological information we share with our interlocutors affect the linguistic features we use? Observations from sign language communities show that there is a relationship between the degree of shared context and lexical variation (e.g., Meir, Israel, Sandler, Padden, & Aronoff, 2012). Iconicity, i.e. form-meaning resemblance, is subjective and depends on one's experience (Occhino, Anible, Wilkinson, & Morford, 2017), and provides the key explanation. Previous theoretical and computational work (Tkachman & Hudson Kam, 2020; Mudd, de Vos, & de Boer, 2022) has proposed that a lack of shared context (i.e., limited context) leads to linguistic alignment because it does not enable the retrieval of meaning from form (i.e., iconicity) when meaning space is not shared. We test this claim by having participants play a communication game about unknown objects in different context conditions. Further, we study what (iconic) strategies individuals align on in the communication game.

Participants (180) recruited on Prolific were assigned to dyads in one of three conditions: a limited context and baseline condition (29 dyads each), and a shared context condition (32 dyads). Stimuli consisted of 12 unfamiliar objects from the NOUN database (Horst & Hout, 2014) and four short videos showing gestural descriptions of each object with different iconic strategies: representing, drawing and two acting strategies (as described in Ortega & Özyürek, 2020). Representing and drawing strategies are iconically motivated by the appearance of the object, while the acting strategy requires understanding an iconic mapping. In phase 1 (training), participants are trained and tested on object-description pairings. In the baseline condition, participants are trained on the same acting description as their partner. In the limited context condition, participants are trained on different acting strategies and in the shared context condition, participants are trained on both acting strategies (see Fig. 1A). In phase 2 (interaction), the dyad plays a communication game over 48 trials where participants alternate as director and matcher. In phase 3 (recall), participants are asked to select the gesture video they would use to successfully communicate with their partner for each object.

We test if dyads in the shared context condition maintain more lexical variation compared to the limited context condition, and find that they do: a linear

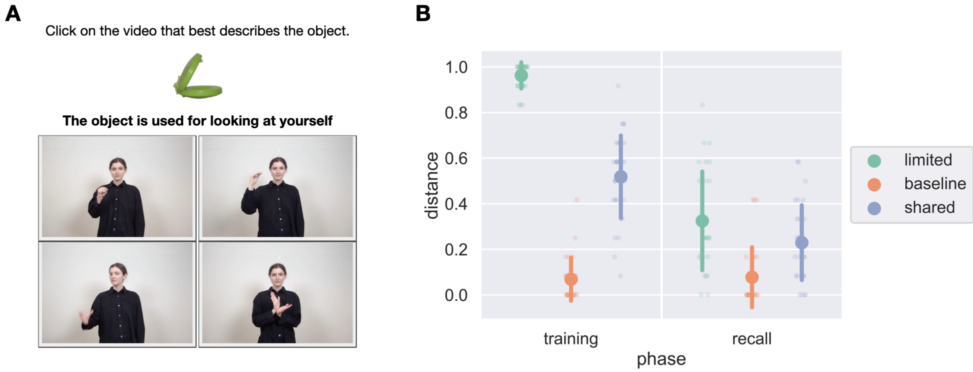


Figure 1. A. An example of experiment phase 1 (training) where participants are trained on object-description pairings. In the limited context condition, one participant would be trained on this acting description ('for looking at yourself') and the other participant would be trained on the other acting description ('for clapping'). In the baseline condition, both participants would be trained on the same acting description. In the shared context condition, participants are trained on both acting descriptions ('for looking at yourself or for clapping'). B. A dot plot showing the mean lexical distance of each condition by phase with error bars showing the standard deviation. Smaller dots show the mean of each dyad. A lexical distance of 0 indicates full alignment. Lexical distance is reduced significantly more over the course of the experiment in the limited context condition compared to the shared context condition ($\chi 2$ (1)=29.55, $p<0.001$).

mixed-effects model with an interaction between condition and phase explains significantly more of the variance than a model without it ($\chi 2$ (1)=29.55, $p<0.001$). However, we cannot exclude the possibility that this effect is an artifact of the high degree of lexical variation in the limited context condition in the training phase, as two additional analyses suggest that in both conditions participants decrease lexical variation to the same extent. The theory suggests that lexical variation decreases due to communication error, so we checked error in interaction across dyads and found the interaction phase of the shared context condition (M=0.06) to be much more comparable to the baseline condition (M=0.04) than to the limited context condition (M=0.22). Focusing on communication strategies, we found interesting differences across conditions; in the limited context condition the proportion of non-acting strategies (drawing and representing) increases significantly from the training phase to the recall phase (β=1.31, SE=0.50, $p<0.001$), likely because their iconicity relates to the physical appearance of the object (shared features for both participants), while in the shared context condition there is no such effect, presumably because participants communicate successfully using the iconic acting strategies, both of which the dyad was trained on. Overall, this work adds support for the link between social structure and linguistic structure (see Lupyan & Dale, 2010).

Acknowledgements

We wish thank the Amsterdam Brain and Cognition research unit at the University of Amsterdam, which has supported Katie with a Talent Grant. Thanks also to the very helpful anonymous reviewers who gave us feedback on our submission!

References

Horst, J., & Hout, M. (2014). *The novel object and unusual name (noun) database: a collection of novel images for use in experimental research.*

Meir, I., Israel, A., Sandler, W., Padden, C. A., & Aronoff, M. (2012). The influence of community on language structure: Evidence from two young sign languages. *Linguistic Variation*, *12*(2), 247–291. (Publisher: John Benjamins)

Mudd, K., de Vos, C., & de Boer, B. (2022). Shared context facilitates lexical variation in sign language emergence. *Languages*, *7*(1), 31.

Occhino, C., Anible, B., Wilkinson, E., & Morford, J. P. (2017). Iconicity is in the eye of the beholder: How language experience affects perceived iconicity. *GEST*, *16*(1), 100–126.

Ortega, G., & Özyürek, A. (2020). Systematic mappings between semantic categories and types of iconic representations in the manual modality: A normed database of silent gesture. *Behav Res*, *52*(1), 51–67.

Tkachman, O., & Hudson Kam, C. L. (2020). Measuring lexical and structural conventionalization in young sign languages. *SL&L*, *23*(1-2), 208–232.

The role of shared labels and shared experiences in representational alignment

Kushin Mukherjee*, Siddharth Suresh*, Xizheng Yu , and Gary Lupyan

Corresponding Author: kmukherjee2@wisc.edu
University of Wisconsin-Madison

Abstract

Successful communication is thought to require members of a speech community to learn common mappings between words and their referents. But if one person's concept of CAR is very different from another person's, successful communication might fail despite the common mappings because different people would mean different things by the same word. Here we investigate the possibility that one source of representational alignment is language itself. We report a series of neural network simulations investigating how representational alignment changes as a function of agents having more or less similar visual experiences (overlap in "visual diet") and how it changes with exposure to category names. We find that agents with more similar visual experiences have greater representational overlap. However, the presence of category labels not only increases representational overlap, but also greatly reduces the importance of having similar visual experiences. The results suggest that ensuring representational alignment may be one of language's evolved functions.

1. Introduction

Imagine two learners of English trying to learn the meanings of "car" and "truck". Some theoretical views describe this process as one of mapping: a word-form is mapped onto the previously existing conceptual categories of CAR and TRUCK (e.g., Fodor, 1975; Snedeker et al., 2004; Pinker, 1994; Bloom, 2002). Alternatively, encountering these labels can help people discover that there is a distinction worth learning and privilege this distinction because it is (apparently) useful in the speech community (Booth & Waxman, 2002; Waxman & Markow, 1995; Xu, 2002; Pomiechowska & Gliga, 2019; Wojcik et al., 2022; Lupyan & Lewis, 2017). In either case, for people to mean similar things by these words would seem to require that the words activate roughly similar semantic representations. For example, if for one

[1] *Equal contribution

person a car is more similar to a truck than to a motorcycle, while for another it is the reverse, we might expect rather severe confusion.

So how do our conceptual representations become aligned such that word meanings can be (more or less) shared? One source of alignment is our shared biology. For instance, the human ear is typically sensitive to a specific range of frequencies and people have roughly similar profiles of sound discrimination (Pumphrey, 1950). This biological commonality ensures that when one person talks about a specific sound (e.g. a sound of a car honking) or tone within this range, another person, barring auditory impairments, will have a similar sensory experience of that sound. Another is shared learning mechanisms - Humans have common cognitive constraints and categorization tendencies that shape how we form concepts. For example, there is little risk of someone's meaning of a 'car' being only blue cars viewed from the side, as this would violate basic principles of human categorization (Rosch & Lloyd, 1978; Shepard, 1994). Another is shared experiences–while each individual's life journey is unique, there are many experiences that are broadly shared. Thus it is possible that human conceptual representations are aligned throughout the process of learning language and it is this alignment that makes linguistic communication possible in the first place. But another possibility is that alignment is achieved—in part—through language itself (Lupyan & Bergen, 2016; Casasanto, 2015; Dingemanse, 2017). On this view, rather than being just a device for conveying our thoughts, language provides an interface between minds (Clark, 1998; Gentner & Goldin-Meadow, 2003; Gomila, Travieso, & Lobo, 2012; Lupyan & Bergen, 2016).

In a prior study, Suffill, van Paridon, and Lupyan (under review) tested the role of language in the conceptual alignment of novel shapes, which could be grouped into two categories based on visual features alone. To test the contributions of verbal labels distinct from perceptual learning towards conceptual alignment, they measured how similarly different participants grouped the shapes in 3 conditions — a baseline condition that relied on the similarity of participants' visual perception, a no-label condition where participants were first familiarized with the category structure of the shapes without labels, and a language condition where they were exposed to incidental nonsense labels for each category. Exposure to labels led to more categorical representations of the concepts (shapes), which in turn led to greater alignment between participants as indicated by more similar sorts.

Here, we build on these findings and prior computational simulations that have hinted at the importance of language in aligning representations of visual concepts (Roads & Love, 2020; Steels, Belpaeme, et al., 2005). Similar to Roads and Love (2020), we explore how learning from multiple signatures of categorical information, feedback from a labeling and a match-to-sample task, affects how the stimuli are represented and the extent to which the representations of different agents (neural networks) are aligned.

Unlike studies with human participants, simulating learning in artificial agents allows us to keep all the learning parameters constant while manipulating the prior perceptual experiences of each agent. Thus, unlike in human behavioral experiments, using neural network model-based agents allows us to examine representational alignment between agents that vary in the overlap of their ‘perceptual diets’ and who are trained with or without category labels.

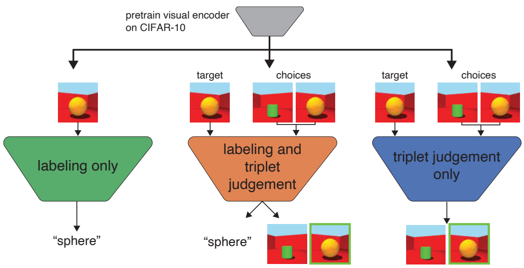


Figure 1. Overview of the 3 task conditions each pre-trained model was fine-tuned on.

1.1. *Dataset*

We leveraged two image databases for out tasks. The first was the CIFAR10 dataset consisting of 60,000 images belonging to 10 object categories. These data was used to pre-train each neural network using the SimCLR unsupervised learning framework (Chen, Kornblith, Norouzi, & Hinton, 2020) so that the networks had some prior visual knowledge (just as human participants do) before being fine-tuned on our experimental dataset and training conditions.

For our main experimental manipulations, we used the Deepmind 3D Shapes dataset (Kim & Mnih, 2018). This dataset consists of 480,000 rendered images of spheres, cubes, cylinders, and capsules varying in size, orientation, and color of the target image and background elements. In our experiments, we kept the color of the background elements constant to simplify learning.

We sampled from this subset of images to create 3 datasets with varying degrees of overlap with each other. Each dataset had 120 images in total with 30 images per shape category randomly sampled from the set of all possible images. We refer to the 3 datasets as **dataset A**, **dataset B**, and **dataset C**. Datasets A and B had 50% overlap in their data, datasets B and C had 33% overlap in their data, and datasets A and C had 0% overlap in their data.

2. Methods

We used `PyTorch` to train neural network models to perform three tasks on the 3D shapes dataset. The three tasks were: (1) labeling the shapes of objects (spheres, cubes, cylinders, or capsules), (2) a match-to-sample triplet similarity judgement task analogous to that used by Suffill et al. (2022)

and (3) a combination of (1) and (2). Models were first pre-trained on the CIFAR10 dataset and were then 'fine-tuned' on the 3D Shapes dataset. Models were trained on one of the three different 3D shape datasets, each of which overlapped with the remaining two to varying degrees. For example, 50% of the images in the first dataset were also present in the second dataset. This allowed us to measure alignment between two models as a function of the overlap in their training as well as whether the training included labels.

2.1. *Model Architecture and Pre-training*

Each model consisted of a simple convolutional encoder consisting of 3 convolutional layers followed by 3 linear 'dense' layers that projected to a 64-dimensional hidden layer. We pre-trained 10 variants of this encoder using the CIFAR10 dataset. Pre-training continued until the validation accuracy was greater than 85% and the mean change in accuracy across epochs was less than 2%. This ensured that all models were trained to a similar criterion before fine-tuning on the 3 task conditions.

2.2. *Training on the Experimental Materials*

We fine-tuned each of the 10 pre-trained models on the 3 tasks below using each of our training datasets — A, B, and C. For each pre-trained model we also fine-tuned a *second* model on dataset A so as to have 2 models that had 100% overlap in training data but different fine-tuning initializations. Thus each of the 10 pretrained models was used to further train 12 models (3 tasks $\times$ 4 datasets). 20 images were held out of each training set and used as a validation set to track network training.

Label condition. In this condition, a 3-layer decoder network took the latent representations from the pre-trained CIFAR10 encoder as input and was tasked with predicting the correct shape label for a given input image. This model was trained on a binary cross-entropy loss on the class logits. Each model was fine-tuned for 1000 epochs, which allowed the validation loss to stabilize.

Triplet Judgement condition. In this condition, the hidden layer of the pre-trained encoder projected to a single linear layer with ReLU activation. We trained the model with a triplet loss objective using the outputs from this layer in the following way. On each iteration, 3 images would be provided to the model — a 'target' image and two 'choice' images. One of the choice images would be exactly identical to the anchor and the other option image would be a random image from one of the three other shape categories. The model's task was to guess which image matched the target image based on the cosine similarities of the latent representations. This model was trained for 1000 epochs, allowing the validation loss to stabilize.

Label and Triplet condition. In this condition, the pre-trained models were tasked with both providing the label for the 'anchor' image as well as performing the triplet judgement task. Both losses were equally weighted and once again the models were trained for 1000 epochs until the validation losses stabilized.

3. Results

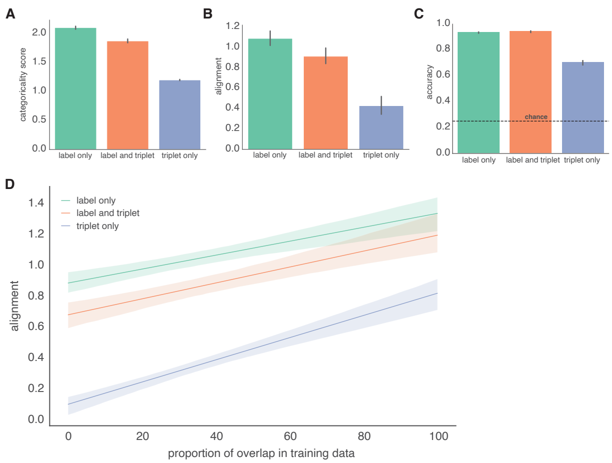


Figure 2. **(A)** Mean categoricality of learned representations, **(B)** mean conceptual alignment between pairs of models in each training condition, and **(C)** mean labeling accuracy in each training condition.**(D)** The effect of data overlap and task on representational alignment.

All results were computed with respect to a set of 480 validation images that were not shown to the networks during training. The validation set consisted of 120 images belonging to each of the 4 shape categories.

Categoricality. The categoricality of the learned representations is the extent to which the networks learned to represent each kind of shape as a distinct category. We quantify categoricality using the activation pattern of the encoder's final hidden layer. Categoricality is defined as follows: $Categoricality = log(\frac{distance_{between-category}}{distance_{within-category}})$, where distance refers to the cosine distance between the activation vectors, and a category consists of, e.g., all the spheres included in the validation set.

Networks trained on only the labeling task showed the greatest of categoricality (M=2.11, sd=.08) followed by networks trained on both labeling and the triplet task (M=1.87, sd=.10). The models trained on only the triplet task showed the least amount of categoricality (M=1.20, sd=.03), all $p's < .001$.

Alignment Alignment is a measure of how similar the representational geometry of a common set of items is across pairs of agents, i.e., neural networks. We operationalized alignment as the log-transformed multiplicative inverse of the Procrustes disparity between the activation vectors from the final encoder layer for the validation images between any given pair of networks. The more similar the representational geometry between the networks the *higher* this alignment value. Pairs of models that were trained on only labeling showed the highest alignment (M= 3.05, sd = .73). Models trained on both tasks showed an intermediate amount of alignment (M= 2.58, sd=.79). Finally, models trained on only the triplet judgement task, i.e., with no category labels, showed the least alignment (M= 1.67, sd=.55), all $p's < .001$. As in the experiment reported by Suffill et al. (under review), categoricality completely mediated the effect of task on alignment. When included as a predictor, the task-associated differences in alignment disappeared ($t's < 1$).

Classification Accuracy We also tested each model on how accurately it could classify the validation images with the correct category label. The triplet-condition models, never trained with labels, could not be expected to produce correct labels and indeed were at chance. To give these models the best possible opportunity to map their learned representations to the correct labels, we fit logistic classifiers using their activation vectors as input and the category labels as output and evaluated using 5-fold cross-validation. We took the mean accuracy on the held-out folds as the labeling accuracy. Networks trained on labels only (M = .93, sd = .01) and labeling and triplet judgements (M = .94, sd = .02) had similarly high performance. Performance of the models trained on only the triplet judgement task was much lower (M = .70, sd = .05), $p < .001$, but well above chance ($p < .001$) showing that it is possible to learn a mapping function from the network's latent states to the labels, albeit not nearly to the same level of accuracy as when the training included labels.

Overlap in 'perceptual diets' To test whether greater amounts of overlap in the training data led to more aligned representations and if this effect varied as function of training task, we fit a linear regression model predicting alignment from the proportion of overlap in training data, the training task, and their interaction. As clearly shown in Figure 2 D., increasing overlap led to greater alignment ($p < .001$). Even with complete overlap in perceptual experience, however, the use of labels continued to have greater alignment. Moreover, decreased overlap impacted alignment between models trained without labels significantly more than either label or label-and-triplet models (p's $< .01$).

4. General Discussion

In a series of simulations we found that training artificial agents on a category learning task with labels led to more categorical representations of concepts relative to a condition with no labels. Additionally, pairs of agents trained with labels showed more conceptual overlap relative to pairs trained without labels. We found that the effect of task on alignment was mediated by the effect of categoricality, which suggests that training with labels induced more categorical representations, which in turn led to greater alignment of agents' representations. In summary, our results highlight the role language might play in aligning our representations of the world so as to facilitate effective communication despite sometimes vast differences in individual experiences (Enfield & Kockelman, 2017).

References

Bloom, P. (2002). *How children learn the meanings of words.* MIT press.

Booth, A. E., & Waxman, S. (2002). Object names and object functions serve as cues to categories for infants. *Developmental psychology*, *38*(6), 948.

Casasanto, D. (2015). All concepts are ad hoc concepts. In *The conceptual mind: New directions in the study of the concepts* (pp. 543–566).

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597–1607).

Clark, A. (1998). Magic words: How language augments human computation. *Language and thought: Interdisciplinary themes*, 162–183.

Dingemanse, M. (2017). On brain-to-brain interfaces, distributed agency and language. In N. J. Enfield & P. Kockelman (Eds.), *Distributed agency* (pp. 59–66). Oxford University Press.

Enfield, N. J., & Kockelman, P. (2017). *Distributed agency.* Oxford University Press.

Fodor, J. (1975). *The Language of Thought.* Cambride, MA: Harvard University Press.

Gentner, D., & Goldin-Meadow, S. (2003). Language in mind: Advances in the study of language and thought.

Gomila, A., Travieso, D., & Lobo, L. (2012). Wherein is human cognition systematic? *Minds and Machines*, *22*, 101–115.

Kim, H., & Mnih, A. (2018). Disentangling by factorising. In *International conference on machine learning* (pp. 2649–2658).

Lupyan, G., & Bergen, B. (2016). How language programs the mind. *Topics in cognitive science*, *8*(2), 408–424.

Lupyan, G., & Lewis, M. (2017). From words-as-mappings to words-as-cues: The role of language in semantic knowledge. *Language, Cognition and Neuroscience*, *34*(10), 1319–1337.

Pinker, S. (1994). *The language instinct. william morrow & co.*

Pomiechowska, B., & Gliga, T. (2019). Lexical acquisition through category matching: 12-month-old infants associate words to visual categories. *Psychological science*, *30*(2), 288–299.

Pumphrey, R. (1950). Upper limit of frequency for human hearing. *Nature*, *166*(4222), 571–571.

Roads, B. D., & Love, B. C. (2020). Learning as the unsupervised alignment of conceptual systems. *Nature Machine Intelligence*, *2*(1), 76–82.

Rosch, E., & Lloyd, B. B. (1978). Principles of categorization.

Shepard, R. N. (1994). Perceptual-cognitive universals as reflections of the world. *Psychonomic Bulletin & Review*, *1*, 2–28.

Snedeker, J., Gleitman, L., et al.. (2004). Why it is hard to label our concepts. *Weaving a lexicon*, *257294*.

Steels, L., Belpaeme, T., et al.. (2005). Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and brain sciences*, *28*(4), 469–488.

Suffill, E., van Paridon, J., & Lupyan, G. (2022). Verbal Labels Increase Conceptual Alignment. In *Joint Conference on Language Evolution* (pp. 691–698).

Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12-to 13-month-old infants. *Cognitive psychology*, *29*(3), 257–302.

Wojcik, E. H., Zettersten, M., & Benitez, V. L. (2022). The map trap: Why and how word learning research should move beyond mapping. *Wiley Interdisciplinary Reviews: Cognitive Science*, *13*(4), e1596.

Xu, F. (2002). The role of language in acquiring object kind concepts in infancy. *Cognition*, *85*(3), 223–250.

Representing space in silent gesture: Communicative contexts compared in Bali

Danielle Naegeli[*1] David Peeters[1], Emiel Krahmer[1], Marieke Schouwstra[2], Made Sri Satyawati[3], and Connie de Vos[1]

[*]Corresponding Author: d.s.u.naegeli@tilburguniversity.edu
[1]Department of Communication and Cognition, Tilburg University, Tilburg, Netherlands
[2]Institute for Logic, Language and Computation, University of Amsterdam, Amsterdam, Netherlands
[3]Fakultas Ilmu Budaya, Universitas Udayana, Denpasar, Bali, Indonesia

Much of our everyday interaction concerns communicating about objects in space. How exactly space gets conceptualised differs between language communities. For instance, in some cultures, like in the Netherlands, space is conceptualised in an egocentric way predominantly using a relative Frame of Reference (FoR). However, we know that other cultures, for instance in Bali, also conceive space geocentrically (Wassman & Dasen, 2006). Being a landmark-based FoR, the four directions used in the Balinese system *kangin, kauh, kelod,* and *kaja* can be interpreted as 'east', 'west', 'seawards', and ‘mountainwards’, the latter referring to the volcano (Gunung Agung) at the centre of Bali (Aryawibawa et al., 2018). While cultures might employ a dominant FoR, this does not mean that they are limited to using just one FoR. Different FoRs can be used depending on situation and addressee, in everyday life of Balinese people, when, e.g., interacting with Indonesians from other islands or foreigners. In this study, we investigated how these FoRs are encoded when Balinese people use silent gesture to convey spatial information.

In addition, the way we conceptualise and the way we (successfully) communicate spatial relations might not be the same across communicative situations. Indeed, within communities what information is conveyed will depend on aspects of the communicative context. For instance, a higher level of communicative demand has been shown to elicit different gestural behaviours (Trujillo et al., 2018). Adding to the complexity of the meaning space in a communicative task by, for example, introducing ambiguity might further push participants to adopt certain strategies and abandon others (cf. Kim & Schachner, 2020).

This preregistered study aimed to investigate the conceptualisation of space as well as the effect of different levels of communicative burden employing the silent gesture paradigm. In silent gesture experiments, hearing participants unfamiliar with communicating in the manual modality were asked to improvise a communication system using only their hands and bodies. Previous research has shown this paradigm to be effective in weakening biases learnt by participants from structure in their spoken language (e.g., Goldin-Meadow et al., 2008). For our study, this also meant participants were unable to rely on spatial terms that correspond to the dominant FoR in their language. Additionally, using silent gesture forced participants to represent their conceptualisation of space within space. This, combined with the different communicative contexts we introduced, allowed us to investigate two core research questions:

1. Removed from spoken language conventions – how do Balinese participants conceptualise and communicate spatial array information?
2. How do participants adjust their strategies according to an increasing communicative burden?

The experiment in our study consisted of three parts of increasing communicative complexity. Each participant was asked to describe 24 stimulus pictures, first, in a non-interactive set-up to a camera (part A), then to an interlocutor in a director-matcher task (part B), and lastly, to the same interlocutor in another director-matcher task in which the meaning space newly included ambiguity (part C). The stimuli were a set of 96 photographs of various Figure-Ground constellations. In the pictures, the Figure object (a toy figurine) was positioned in four possible locations and with four possible orientations around the inanimate Ground object (inspired by a paradigm established by Levinson et al., 1992).

We have concluded the data collection from 24 participants at our Balinese fieldwork site at Universitas Udayana in Denpasar, Bali. Preliminary analysis of the Balinese data revealed that participants employed predominantly either mirror-image depiction or absolute translation of the spatial constellation of the stimulus pictures. Furthermore, we found that in part A, the single-participant non-interactive set-up, participants produced a high number of incomplete depictions that would not be sufficient to recognise the target picture. This number of incomplete depictions decreased rapidly for the interactive director-matcher task part B and even further for part C. Theoretical implications of these findings will be discussed.

References

Aryawibawa, I. N., Putra Yadnya, I. B., Ngurah Parthama, I. G., & Pye, C. L. (2018). Balinese spatial reference frames: Linguistic and non-linguistic evidence from the north of Bali. *Lingua, 215*, 40–52. https://doi.org/10.1016/j.lingua.2018.09.005.

Goldin-Meadow, S., So, W. C., Ozyürek, A., & 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.

Kim, M., & Schachner, A. (2020). Gesture Production and Theory of Mind: Effective Disambiguation in Communication through Gesture. https://doi.org/10.31234/osf.io/xdsvj

Levinson, S. C., Brown, P., Danzinger, E., De León, L., Haviland, J. B., Pederson, E., & Senft, G. (1992). Man and tree & space games. In *Space stimuli kit 1.2* (pp. 7-14). Max Planck Institute for Psycholinguistics.

Trujillo, J. P., Simanova, I., Bekkering, H., & Ozyürek, A. (2018). Communicative intent modulates production and comprehension of actions and gestures: A Kinect study. *Cognition, 180*, 38–51. https://doi.org/10.1016/j.cognition.2018.04.003.

Wassmann, J., & 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* (pp. 351-376). Bielefeld: Transcript Publishers.

Form and function in the evolution of symbolic artifacts: A transmission chain study

Murillo Pagnotta[1,*], Izzy Wisher[1,2,], Malte Lau Petersen[1], Felix Riede[2], Riccardo Fusaroli[1], Kristian Tylén[1]

*Corresponding Author: murillo@cc.au.dk
[1] Department of Linguistics, Cognitive Science and Semiotics, Aarhus University, Denmark
[2] Department of Archaeology and Heritage Studies, Aarhus University, Denmark

1. Introduction and Methods

Capacities for symbolic behavior—here understood in the broadest sense to include practices of decoration, marking of identity, and communication—are considered constitutive of the human species and permeate almost all aspects of contemporary human life (Deacon, 1998; Donald, 1991; Pagnotta, 2014). Yet, we still lack a detailed understanding of the mechanisms by which such capacities evolved during the Late Pleistocene. Intentional markings left on rock and bone surfaces are an important source of evidence, but these are often opaque with respect to their original meaning and function (Overmann & Coolidge, 2019).
The engraved markings on ochre and ostrich eggshell from the Blombos Cave and the Diepkloof Rock Shelter, South Africa, date back to a prolonged period from 100.000 to 60.000 BP (Henshilwood et al., 2009; Texier et al., 2010). There have been heated controversies about the past function of the engravings, which have been proposed to form part of an aesthetic activity (Hodgson, 2014), an expression of group identity (Texier et al., 2013), or a medium of denotational communication (Henshilwood et al., 2009). Interestingly, the composition of engravings shows profound structural changes over time with early patterns consisting of simple parallel lines while later patterns become increasingly regular and ordered with oblique and hashtag-like intersecting lines (fig. 1).

Figure 1. Examples of the Blombos ochre engravings dated to the period c. 100.000 - 70.000 BP, organized with older items to the left and younger to the right.

Here we used outlines of the Blombos and Diepkloof engravings and combined transmission chains and perceptual experiments to systematically examine whether markings evolve different cognitive implications contingent on their immersion in different contexts of use. First, we had 120 participants reproduce and transmit stimuli derived from the oldest engraved patterns at Blombos and Diepkloof over eight generations in three conditions, as part of either i) a decorative, ii) identity marking, or iii) communicative activity. We then used five perceptual experiments to examine the cognitive implications of changes accumulated in the markings over generations. Specifically, we examined whether markings became i) more salient to the human eye; ii) more likely to be recognized as purposefully made by another human; iii) easier to reproduce from memory; iv) easier to recognize as originating from a specific group; and v) more discriminable from each other. Based on Tylén et al (2020), we hypothesized that the markings produced under the different conditions of the transmission chains would show different profiles of change with implications for perception and cognition (see predictions in Table 1).

Table 1. Predicted direction of changes over generations

	Saliency	*Intentionality*	*Memorability*	*Style*	*Discriminability*
Decoration	increase	increase	no change	no change	no change
Group identity	increase	increase	increase	increase	no change
Communication	increase	increase	increase	no change	increase

2. Results and Discussion

Preliminary analyses indicate that markings produced within the contexts of different symbolic practices indeed evolve structural properties that differentially engage human cognitive processes (Fig. 2). And by comparing the condition-related profiles of change over generations to the evolution of the original Blombos and Diepkloof engravings, we demonstrate how we can inform inferences about their past symbolic function.

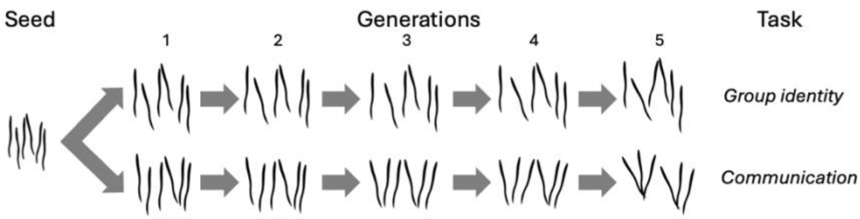


Figure 2. Examples of patterns produced by two independent transmission chains. The same initial seed stimulus accumulates different structural changes depending on the reproduction context.

Acknowledgements

The authors acknowledge funding from an ERC Consolidator Grant (101044626 - eSYMb) awarded to Kristian Tylén.

References

Deacon, T. W. (1998). *The symbolic species: The co-evolution of language and the brain*. WW Norton & Company.

Donald, M. (1991). *Origins of the Modern Mind: Three Stages in the Evolution of Culture and Cognition*. Harvard University Press.

Henshilwood, C. S., d'Errico, F., & 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*(01), 57–69.

Overmann, K. A., & Coolidge, F. L. (2019). *Squeezing minds from stones: Cognitive archaeology and the evolution of the human mind*. Oxford University Press.

Pagnotta, M. (2014). On the controversy over non-human culture: The reasons for disagreement and possible directions toward consensus. *Behavioural Processes*, *109*, 95–100. https://doi.org/10.1016/j.beproc.2014.04.008

Texier, P.-J., Porraz, G., Parkington, J., Rigaud, J. P., Poggenpoel, C., Miller, C., Tribolo, C., Cartwright, C., Coudenneau, A., Klein, R., Steele, T., & Verna, C. (2010). A Howiesons Poort tradition of engraving ostrich eggshell containers dated to 60,000 years ago at Diepkloof Rock Shelter, South Africa. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(14), 6180–6185. https://doi.org/10.1073/pnas.0913047107

Texier, P.-J., Porraz, G., Parkington, J., Rigaud, J.-P., Poggenpoel, C., & Tribolo, C. (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.

Tylén, K., Fusaroli, R., Rojo, S., Heimann, K., Fay, N., Johannsen, N. N., Riede, F., & Lombard, M. (2020). The evolution of early symbolic behavior in Homo sapiens. *Proceedings of the National Academy of Sciences*, *117*(9), 4578–4584.

The Evolution of Reference

Giulia Palazzolo[1,2]

*Corresponding Author: giulia.palazzolo.1@warwick.ac.uk
[1]Department of Philosophy, University of Warwick, Coventry, UK
[2]Department of Philosophy, University La Sapienza, Rome, Italy

Abstract

An important question in the studies of language evolution is whether there are traces of reference in animal signaling behaviors; or if, by contrast, reference is a unique feature of human language. After Seyfarth et al. (1980)'s study of vervet monkeys alarm calls, three main theories have come up in the literature, which seek to characterize animal signaling behaviors as evolutionary precursors of linguistic reference. These are: the theory of functional reference (Marler, Evans & Hauser, 1992), the meaning attribution framework (Wheeler & Fischer, 2012), and the revised version of functional reference (Scarantino, 2013).

In this talk, I will start by examining the different ways in which the theory of functional reference, the meaning attribution framework, and the revised version of functional reference conceptualize animal reference. Then, I will turn to spelling out some limits that these frameworks encounter as accounts of the evolution of linguistic reference. I will argue that functional reference can be advantageous when studying animal communication systems in their own right. As a functional framework, it groups together signals that achieve a similar referential function, allowing exploration into the diverse ways in which animals can provide recipients with information about objects. However, when the goal shifts to pinpointing the evolutionary precursors of linguistic reference, functional reference becomes less advantageous. Since it is neutral about what mechanisms underpin signal production, functional reference incorporates phenomena that aren't plausibly evolutionarily connected to linguistic reference, because they are underpinned by substantially different mechanisms. They are more likely analogous than homologous (see e.g., Scott-Phillips & Heintz, 2023).

Regarding Wheeler and Fischer's framework, I will additionally argue that their framework appears incapable of isolating phenomena that are directly relevant to an account of the evolution of reference. This is because, while the

meaning-attribution framework centers on the receiver, reference primarily constitutes an act performed by the producer.

I will show that, despite the differences between these frameworks, there exists a common thread binding them together: the idea that the mechanisms of signal production in animals are essentially different from those of humans, and thus uninteresting for understanding the evolution of linguistic reference – an intentional act of drawing someone's attention to an object (see e.g., Bach, 2008). Based on this premise, these theoretical frameworks set out to explore the evolution of linguistic reference while maintaining a neutral stance on signal production.

In the second part of the talk, I will refer to a recent study by Crockford et al. (2012, 2017) on chimpanzees' alert hoos, suggesting that at least some animal communicative acts display strong psychological parallels with linguistic reference. I will argue that, from an evolutionary viewpoint, the fact that there are signaling behaviors in chimps are psychologically similar to linguistic reference has some important implications.

Firstly, this psychological continuity leads us to consider the possibility that the mechanisms for reference may be homologous in chimps and humans (i.e., might share a common evolutionary origin). Thus, that the intentional act of drawing someone's attention to an object (i.e., human-like intentional reference) may not be an exclusive human trait, but a capacity that was present in our last common ancestor (LCA). This hypothesis is reinforced, among other things, by similar findings in bonobos (Girard-Buttoz et al., 2020). However, it is important to clarify that this claim about homology does not necessarily extend to other features of the producer's psychology in communication (cf. Moore, 2017; Bar-On, 2021; Warren & Call, 2022; Scott-Phillips & Heintz, 2023).

Secondly, the existence of animal intentional reference grants us the opportunity to go beyond functional reference and meaning attribution in the study of the evolution of reference. Drawing on the works of Crockford et al. (2012, 2017), and Girard-Buttoz et al. (2020), I will propose a novel account for the study of the evolution of linguistic reference, which focuses on the mechanisms of signal production, introducing the following constraint: (i) an utterer produces a signal with the intention to direct a receiver's attention to an object. I will show that this account is more suitable than existing frameworks (i.e., functional reference and meaning attribution) when it comes to identifying potential evolutionary precursors of linguistic reference in animal communication. Looking at the evolutionary phylogeny of our capacity for reference, I will argue, holds significant value for understanding its distinctive properties, selective advantages, and specific evolutionary history.

References

Bach, K. (2008). What does it take to refer?. In E. Lepore, B.C. Smith (Eds.), *The Oxford Handbook of Philosophy of Language* (pp. 516-554). Oxford: Oxford UP.

Bar-On, D. (2021). How to Do Things with Nonwords: Pragmatics, Biosemantics, and Origins of Language in Animal Communication. *Biology & Philosophy*, *36*(6), 50.

Crockford, C., Wittig, R.M., Mundry, R., & Zuberbühler, K. (2012). Wild Chimpanzees Inform Ignorant Group Members of Danger. *Current Biology*, *22*(2), 142-146.

Crockford, C., Wittig, R.M., & Zuberbühler, K. (2017). Vocalizing in Chimpanzees Is Influenced by Social-Cognitive Processes. *Science Advances*, *3*(11), e1701742.

Girard-Buttoz, C., et al. (2020). Information transfer efficiency differs in wild chimpanzees and bonobos but not social cognition. *Proceedings Royal Society B*, *287*, 20200523.

Marler, P., Evans, C.S., & Hauser, M.D. (1992). Animal Signals: Motivational, Referential, or Both? In H. Papousek, U. Jürgens (Eds.), *Nonverbal Vocal Communication: Comparative and Developmental Approaches* (pp. 66-86). Cambridge: Cambridge University Press.

Moore, R. (2017). Gricean Communication and Cognitive Development. *Philosophical Quarterly*, *67*(267), 1-24.

Scarantino, A. (2013). Rethinking Functional Reference. *Philosophy of Science*, *80*(5), 1006-1018.

Scott-Phillips, T.C., & Heintz, C. (2023). Animal Communication in Linguistic and Cognitive Perspective. *Annual Review of Linguistics*, *9*, 93-111.

Seyfarth, R.M., Cheney, D.L., & Marler, P. (1980). Monkey Responses to Three Different Alarm Calls: Evidence of Predator Classification and Semantic Communication. *Science*, *210*(4471), 801-803.

Warren, E., & Call, J. (2022). Inferential Communication: Bridging the Gap Between Intentional and Ostensive Communication in Non-Human Primates. *Frontiers in Psychology*, *12*, 718251.

Wheeler, B.C., & Fischer, J. (2012). Functionally Referential Signals: A Promising Paradigm Whose Time Has Passed. *Evolutionary Anthropology: Issues, News, and Reviews*, *21*(5), 195-205.

Why are languages skewed? A Bayesian account for how skew and type count, but not entropy, facilitate rule generalisation

Elizabeth Pankratz[*1], Simon Kirby[1], and Jennifer Culbertson[1]

[*]Corresponding Author: e.c.pankratz@sms.ed.ac.uk
[1]Centre for Language Evolution, University of Edinburgh, Edinburgh, UK

One reason that languages come to have the properties they do is because those properties make a language easier to learn, its rules easier to generalise (Kirby, Griffiths, & Smith, 2014). Here, we ask which statistical properties of a language help people to decide whether a linguistic rule can be extended to new instances.

Previous research has generated seemingly conflicting hypotheses regarding the role of *Shannon entropy*. Segmentation and generalisation (closely related processes; Frost & Monaghan, 2016) are facilitated when items the rule applies to follow a skewed frequency distribution (e.g., Kurumada, Meylan, & Frank, 2013; Lavi-Rotbain & Arnon, 2019b, 2019a, 2021, 2022; Casenhiser & Goldberg, 2005; Goldberg, Casenhiser, & Sethuraman, 2004). Since skewed distributions have lower entropy than uniform distributions over the same number of items, this finding has been explained as a facilitatory effect of low entropy (e.g., Lavi-Rotbain & Arnon, 2022). At the same time, rules are more readily generalised when they apply to a large number of distinct types (e.g., Gómez, 2002; Tamminen, Davis, & Rastle, 2015; Valian & Coulson, 1988; Radulescu, Wijnen, & Avrutin, 2020). This result has also been explained in terms of entropy—but now, since a distribution over more types has higher entropy than a distribution over fewer, the prediction is that *high* entropy prompts generalisation. How do these seemingly contradictory findings fit together?

In this preregistered artificial language learning experiment (`osf.io/5keh9`), we disentangle how skew, type count, and entropy influence generalisation. Participants learned two different plural suffixes, each occurring with stems that followed one of two frequency distributions (Figure 1A). Then at test, they were asked to choose which of the two suffixes to use with novel stems. For half of the participants, the distributions that were contrasted were a uniform distribution over four types (Unif4) vs. a skewed distribution over four types (Skew4). For the other half, the distributions were, again, Unif4 vs. a uniform distribution over twice as many types (Unif8). Including Unif4 in both groups gives a baseline from which we can evaluate the individual effects of skew (in Group 1) and type count (in Group 2).

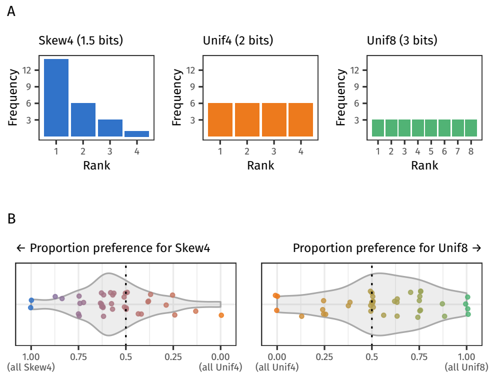


Figure 1. (A) The frequency distributions compared (Group 1 saw Skew4 and Unif4; Group 2 saw Unif4 and Unif8). (B) The suffixes that perfect learners generalised with tend to be the ones that appear with more variable stems (Skew4 in Group 1; Unif8 in Group 2). (C) This finding aligns with the posterior probabilities of missing a type when sampling from posited latent distributions.

We analyse data for participants who perfectly learned the language (N = 77 of 100, split 38–39 between groups). We found that Group 1 preferred to generalise with Skew4, while Group 2 preferred Unif8 (Figure 1B). A Bayesian linear model estimated the probability of generalising with the non-baseline suffix as 56.3% (95% CrI: [49.8%, 62.8%]). The same model estimated no difference between the two participant groups (β = 0 log-odds, 95% CrI [–0.49, 0.51]); skew and a greater type count provide comparable evidence that the suffix can be generalised.

These results are consistent with the empirical findings of both sets of studies summarised above, and thus *not* with an explanation based on entropy. We propose instead that the explanation follows from participants reasoning in a probabilistic, Bayesian way. In particular, participants in our task must essentially guess which suffix is more likely to appear with additional types beyond the ones they've already seen. If Unif4 were only a sample from some larger latent distribution, say Unif5, then observing only Unif4 and missing that fifth type is relatively unlikely. But missing a type when sampling from a latent Skew5 or Unif9 would be much more likely because of the greater number of low-frequency types overall.

Figure 1C shows each group's posterior probabilities of failing to encounter one or more types when sampling from each latent distribution. These probabilities heavily favour the distribution that learners in our experiment preferred. The greater generalisability of rules with these features could be part of the explanation for why languages come to have properties such as skew; ultimately, probabilistic reasoning of the kind we observe may shape the statistical structure of language.

Acknowledgements

EP gratefully acknowledges funding from the Economic and Social Research Council (award number ES/P000681/1) and the Social Sciences and Humanities Research Council of Canada (award number 752-2021-0366).

References

Casenhiser, D. M., & Goldberg, A. E. (2005). Fast mapping of a phrasal form and meaning. *Developmental Science*, *8*(6), 500–508.

Frost, R. L., & Monaghan, P. (2016). Simultaneous segmentation and generalisation of non-adjacent dependencies from continuous speech. *Cognition*, *147*, 70–74.

Goldberg, A. E., Casenhiser, D. M., & Sethuraman, N. (2004). Learning argument structure generalizations. *Cognitive Linguistics*, *15*(3).

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*(5), 6.

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, *28*, 108–114.

Kurumada, C., Meylan, S. C., & Frank, M. C. (2013). Zipfian frequency distributions facilitate word segmentation in context. *Cognition*, *127*(3), 439–453.

Lavi-Rotbain, O., & Arnon, I. (2019a). Children learn words better in low entropy. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (pp. 631–637).

Lavi-Rotbain, O., & Arnon, I. (2019b). Low entropy facilitates word segmentation in adult learners. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (pp. 2092–2097).

Lavi-Rotbain, O., & Arnon, I. (2021). Visual statistical learning is facilitated in Zipfian distributions. *Cognition*, *206*, 104492.

Lavi-Rotbain, O., & Arnon, I. (2022). The learnability consequences of Zipfian distributions in language. *Cognition*, *223*, 105038.

Radulescu, S., Wijnen, F., & Avrutin, S. (2020). Patterns bit by bit: An entropy model for rule induction. *Language Learning and Development*, *16*(2), 109–140.

Tamminen, J., Davis, M. H., & Rastle, K. (2015). From specific examples to general knowledge in language learning. *Cognitive Psychology*, *79*, 1–39.

Valian, V., & Coulson, S. (1988). Anchor points in language learning: The role of marker frequency. *Journal of Memory and Language*, *27*(1), 71–86.

Do semiotics experiments really show the "superiority" of gesture over vocalization for iconic representation? And even if they do, does it matter?

Marcus Perlman[*1] and Vinicius Macuch Silva[1]

[*]Corresponding Author: m.perlman@bham.ac.uk
[1]English Language and Linguistics, University of Birmingham, Birmingham, United Kingdom

Sometime in the emergence of language, our ancestors faced the challenge of creating new symbols when there were none before. A version of the symbol grounding problem, this ability to establish meaningful symbols without convention is noted as a major hurdle that needed to be overcome to get language off the ground. This has led some researchers to argue for a gesture-first origin of language, based largely on the premise that gestures afford vastly more iconicity than what is assumed to be the only negligible iconicity afforded by vocalizations (e.g., Arbib, 2005; Armstrong & Wilcox, 2007; Corballis, 2003; Sandler, 2013). Thus the use of gestures, but not vocalizations, would have enabled our ancestors to create "already meaningful" signals that could, in turn, be conventionalized into symbols (Tomasello, 2008). Then, at a later stage, vocalizations, intrinsically void of iconicity, would have needed to piggy-back off meaningful gestures – perhaps even fully-fledged signed languages – to bootstrap the transition to the first spoken languages.

A major source of evidence indicating the so-called superiority of gestures over vocalizations for iconic representation comes from semiotics experiments in which participants play "charades" using either gestures or vocalizations, allowing – in theory – a semiotic comparison between them (Fay et al., 2013, 2014, 2022). These experiments have generally found that participants communicate more accurately with gestures than vocalizations, which is taken as evidence of gesture's greater iconic potential. In this paper, we argue that these experiments, while informative, are fundamentally limited in ways that make the task of comparing gestures and vocalizations unbalanced if not impossible. We provide both methodological and theoretical reasons for why such experiments put vocalizations on unequal footing with gestures, biasing the results towards the conclusion that gestures are superior. First, we explain how these semiotics experiments, by their design, constrain the use of convention differently in each modality. Producers must suppress iconic words in the vocal condition, whereas they are permitted to gesture freely in the gesture condition – a critical difference that confounds the comparison between the modalities. Second, we argue that these experiments, by implementing an unnatural division that equates vocalization with purely acoustic communication and gesture with purely visual communication, pits the modalities against each other in a completely unrealistic scenario that favours gesture.

By highlighting these issues, we hope to inform future semiotics experiments seeking to compare different modalities in their potential for iconic communication, and concomitantly, for grounding new symbols. We conclude by questioning whether it even matters if gestures are “superior” to vocalizations in their potential for iconicity. While guessing accuracy may be, on average, higher with gestures, semiotics experiments show that vocalizations also afford plenty of iconicity (Ćwiek et al., 2021), which may be sufficient to ground the formation of vocal symbols (Perlman et al., 2015). This point raises important questions for understanding the *complementary* roles of vocalization and gesture in a multimodal evolution of language (Macuch Silva et al., 2020).

References

Arbib, M. A. (2005). From monkey-like action recognition to human language: An evolutionary framework for neurolinguistics. *Behavioral and Brain Sciences*, *28*(02), 105–124.

Armstrong, D. F., & Wilcox, S. E. (2007). *The Gestural Origin of Language*. Oxford University Press.

Corballis, M. C. (2003). *From hand to mouth: The origins of language*. Princeton University Press.

Ćwiek, A., Fuchs, S., Draxler, C., Asu, E. L., Dediu, D., Hiovain, K., Kawahara, S., Koutalidis, S., Krifka, M., Lippus, P., Lupyan, G., Oh, G. E., Paul, J., Petrone, C., Ridouane, R., Reiter, S., Schümchen, N., Szalontai, Á., Ünal-Logacev, Ö., … Perlman, M. (2021). Novel vocalizations are understood across cultures. *Scientific Reports*, *11*(1), Article 1. https://doi.org/10.1038/s41598-021-89445-4

Fay, N., Arbib, M., & 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., Lister, C. J., Ellison, T. M., & Goldin-Meadow, S. (2014). Creating a communication system from scratch: Gesture beats vocalization hands down. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00354

Fay, N., Walker, B., Ellison, T. M., Blundell, Z., De Kleine, N., Garde, M., Lister, C. J., & Goldin-Meadow, S. (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

Macuch Silva, V., Holler, J., Ozyurek, A., & 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

Perlman, M., Dale, R., & 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

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

Tomasello, M. (2008). *Origins of human communication*. MIT press.

Does syntactic alignment predict cooperation? A corpus study of the prisoner's dilemma

Marek Placiński[*1], Theresa Matzinger[2,3], Andreas Baumann[4], Przemysław Żywiczyński[1], Stefan Hartmann[4], Irene Boehm[2], Michael Pleyer[1], and Sławomir Wacewicz[1]

[*]Corresponding Author: marpla@umk.pl
[1]Center for Language Evolution Studies, Nicolaus Copernicus University in Toruń, Toruń, Poland
[2]Department of English and American Studies, University of Vienna, Vienna, Austria
[3]Cognitive Science Hub, University of Vienna, Vienna, Austria
[4]Department of German Studies, University of Vienna, Vienna, Austria
[4]German Department, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

1. Introduction

The cooperative nature of human communication is widely accepted as a fact at least since Grice (1975; but also Clark, 1996). In this study, we investigate whether structural alignment correlates with cooperative behaviour. Some evidence suggests that structural alignment is sensitive to extralinguistic factors such as power (Danescu-Niculescu-Mizil et al., 2012) or status (Lev-Ari & Peperkamp, 2017). However, whether structural alignment itself increases the propensity to cooperate remains unexplored, although some studies point to the fact that lexical imitation leads to prosocial behaviour (van Baaren et al., 2004). Here, we aim to test whether structural alignment affects decisions in a cooperative task.

2. Data

This study uses text transcripts from *Golden Balls* (2007–2009), a TV show in which four contestants play four rounds of a game, voting out one player until only two remain. In the final round of the game, the two contestants can either split the jackpot (divide evenly) or steal it (claim for oneself) at the end of the game. Mutual splitting is better than mutual stealing (each contestant receives half of the jackpot vs nothing), but stealing while the other participant splits ensures the biggest payoff for the defecting contestant. This payoff structure makes the game formally equivalent to the Prisoner's Dilemma, the game format traditionally used in behavioral economics to model cooperation (Rapoport, 1989). Importantly for the aim of this study, the contestants make their decision based solely on previous interaction with each other.

3. Method and results

Seventeen *Golden Balls* transcripts were parsed for constituency structure with the use of CoreNLP probabilistic context free grammar parser (Manning et al., 2014). These trees were subsequently transformed into production rules (of the form NP → Det N). Unary and lexical productions were removed from the dataset. In total, we obtained 71078 productions. The productions were then automatically annotated for repetitions, and any repetition of a production rule was considered a case of syntactic alignment. Repetitions arising from lexical overlaps were removed from the analysis. We controlled for pre-established linguistic similarity by applying a sliding window and considering the span of 50 previous productions. To test our hypothesis, we fit a GLMM model with alignment as the predictor and contestants' decision (split vs steal) as the outcome variable. We found a positive relation between alignment and cooperation ($\beta = 0.05$, $p = 0.03$).

4. Discussion

The results of our study suggest that syntactic alignment correlates with cooperation in the real world. This is consistent with the interpretation that structural alignment may be a sort of low-level signal/cue that truthfully informs about an individual's disposition to cooperate (Wacewicz et al., 2017) – possibly because structural alignment is difficult to fake (Bargh & Chartrand, 1999). Alternatively, alignment might convey a degree of similarity with others, which has also been shown to promote cooperation (McNeill, 1995).

Acknowledgements

Marek Placiński was financially supported by the Polish National Science Centre [grant agreement number UMO-2019/33/N/HS2/00541].
Michael Pleyer was supported by project No. 2021/43/P/HS2/02729 co-funded by the National Science Centre and the European Union Framework Programme for Research and Innovation Horizon 2020 under the Marie Skłodowska-Curie grant agreement No. 945339.

References

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception–behavior link and social interaction. *Journal of Personality and Social Psychology, 76*(6), 893–910.

Clark, H. H. *Using Language.* Cambridge: Cambridge University Press.

Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., & Kleinberg, J. (2012). Echoes of Power... In *Proc 21st Int. Conference on World Wide Web* (pp. 699–708). Association for Computing Machinery.

Grice, H. P. (1975). Logic and conversation. In P. Cole, & J. Morgan (Eds.), *Syntax and Semantics* (pp. 41-58). New York: Academic Press.

Lev‐Ari, S., Heugten, M., & Peperkamp, S. (2017). Relative difficulty of understanding foreign accents as a marker of proficiency. *Cognitive Science, 41*(4), 1106–1118.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55–60.

McNeill, W. H. (1995). *Keeping Together in Time: Dance and Drill in Human History*. Harvard University Press. https://doi.org/10.2307/j.ctvjf9wq6

Rapoport, A. (1989). Prisoner's Dilemma. In: Eatwell, J., Milgate, M., Newman, P. (eds) *Game Theory* (pp. 199–204). Palgrave Macmillan, London. https://doi.org/10.1007/978-1-349-20181-5_23

Wacewicz S., Żywiczyński P., Chiera, A. (2017). An evolutionary approach to low-level conversational cooperation. *Language Sciences 63*, 91–104. https://doi.org/10.1016/j.langsci.2017.01.005

van Baaren, R. B., Holland, R. W., Kawakami, K., & van Knippenberg, A. (2004). Mimicry and Prosocial Behavior. Psychological Science, 15(1), 71–74. https://doi.org/10.1111/j.0963-7214.2004.01501012.x

Towards computational detection of metaphoric change in language evolution via word embeddings

Michael Pleyer[*1], Klaudia Karkowska[1], Svetlana Kuleshova[1,2], Darya Namednikava[1], and Marek Placiński[1]

[*]Corresponding Author: pleyer@umk.pl
[1]Center for Language Evolution Studies, Nicolaus Copernicus University in Toruń, Toruń, Poland
[2]ArScAn-Équipe AnTET (UMR 7041), Université Paris Nanterre, Nanterre, France

1. Introduction

One of the key communicative challenges that the hominin lineage was faced with in the course of the evolution of language was how to communicate about new referents in a changing, dynamic environment that continually introduced new potential referents and affordances of things to communicate about, while only possessing a limited inventory of signals. One of the solutions to this problem observable in the cultural evolution of languages is the process of meaning extension, whereby existing words acquire additional meanings, thereby becoming polysemous (e.g. Srinivasan & Rabagliati 2015). One of the primary driving factors of such semantic change is that of metaphor (Anderson 2017). For this reason, metaphor has also been assigned a central role in the evolution of language (e.g. Smith & Höfler 2015; Ellison & Reinöhl 2022).

Here, we propose a computational method to further investigate the processes of metaphoric extension central to language evolution. Computational methods have shown increasing promise in automatically detecting metaphoric change in language (e.g. Schlechtweg et al. 2017; Hamilton et al., 2016; Hills & Miani in press; Lau et al., 2012). However, one of the challenges of these approaches is that they generally require large corpus sizes. Using a less-resourced language, Polish, as a case study, we show how some of these challenges might potentially be addressed. Specifically, we explore the potential of using word embeddings in detecting metaphorical change in technology-related expressions in Polish.

2. Methods

This study analyses 60 words related to science and technology from the Korpus Barokowy (Kieraś et al., 2017; *KorBa*), a corpus of 17–18th century Polish and the corpus of 19th and early 20th century Polish (Łaziński et al., 2023, *F19*). Metaphors from these source domains emerged relatively recently, in contrast with figurative expressions related to natural phenomena (such as *fire* and *cold*), which opens a greater possibility for detecting the acquisition of metaphorical meanings in historical corpora.

In the first step of the analysis, we extracted all sentences containing the words from the list. The F19 dataset contained 1744 sentences, whereas the KorBa dataset included 1210. These sentences subsequently underwent binary annotation for whether the use of the target word in that sentence was metaphorical or not. Afterwards, we extracted sentence-level embeddings and word embeddings for the target word by fine-tuning the transformer-based large language model Polbert (Kłeczek et al., 2020), a Polish version of the BERT language mode (Devlin et al. 2019), to both of our datasets. We then computed the cosine distance between the word embedding and the sentence-level embedding, hypothesising that greater cosine distance entails a greater probability of that word being metaphorical (Liu et al., 2020).

For the classification task, both datasets were split so that 80% of the dataset was used as the training data and the remaining 20% was used for testing. We used a logistic regression classifier to evaluate our hypothesis. Our model achieved an F1-score of 60% for the KorBa corpus and an F1-score of 68% for the F19 corpus. We then extracted words that were classified as non-metaphorical in the KorBa corpus and compared this list with the list of metaphorical words from the F19 corpus.

One of the examples is the word *komórka* (*cell*), which occurs in contexts related to homes in KorBa (the word is a diminutive of *komora – chamber*), and acquires the metaphorical meaning of a biological cell in PL19. A similar case is attested for the word *ropa* (*puss*, but also *oil*), where the former meaning is attested in KorBa, but both meanings in PL19.

3. Conclusion

Given the proposed importance of metaphoric change for language evolution, it is important to investigate its dynamics in observable language change. Here we show that using fine-tuned state-of-the-art language models to historical corpora can support analyses of the acquisition of metaphorical meaning.

Acknowledgements

This research is part of the project No. 2021/43/P/HS2/02729 co-funded by the National Science Centre and the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 945339. For the purpose of Open Access, the author has applied a CC-BY public copyright licence to any Author Accepted Manuscript (AAM) version arising from this submission

References

Anderson, W. (2017). Metaphor and diachronic variation. In E. Semino & Z. Demjén (Eds.), *Routledge Handbook of Metaphor and Language* (pp. 233-246). Routledge.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of NAACL-HLT 2019* (pp. 4171-4186).

Ellison, T. M., & Reinöhl, U. (2022). Compositionality, metaphor, and the evolution of language. *International Journal of Primatology*, 1-17.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. doi.org/10.18653/v1/P16-1141

Hills, Thomas & Alessandro Miani. In press. A short primer on historical natural language processing. In Ganna. Progrebna & Thomas Hills (eds.) Handbook of Behavioural Data Science. Cambridge University Press.

Jurafsky, D., & Martin, J. H. (2023). Speech and Language Processing. Prentice Hall PTR.

Kieraś W., Komosińska D., Modrzejewski E., Woliński M. (2017). Morphosyntactic annotation of historical texts. The making of the baroque corpus of polish. Text, Speech, and Dialogue 20th International Conference, 308–316.

Kłeczek, D. (2020). Polbert: Attacking Polish NLP Tasks with Transformers. In *Proceedings of the PolEval 2020 Workshop* (pp. 79–88).

Lau, J. H., Cook, P., McCarthy, D., Newman, D., & Baldwin, T. (2012). Word sense induction for novel sense detection. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 591–601.

Liu, J., O'Hara, N., Rubin, A., Draelos, R., Rudin, C. (2020). Metaphor detection using contextual word embeddings from transformers. In *Proceedings of the Second Workshop on Figurative Language Processing* (pp. 250-255).

Łaziński, M., Górski, R. L. i Woźniak, M. (2023). Korpus XIX w. Uniwersytetu Warszawskiego i IJP PAN. LingVaria, 18(1). doi: 10.12797/LV.18.2023.35.09.

Schlechtweg, D., Eckmann, S., Santus, E., Schulte im Walde, S. & Hole, D. 2017. German in Flux: Detecting Metaphoric Change via Word Entropy. *Proceedings of the 21st Conference on Computational Natural Language Learning* (CoNLL 2017), 354-357. https://doi.org/10.18653/v1/K17-1036

Smith, A. D., & Höfler, S. (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: Perspectives on the Sociohistorical Linguistics of Figurative Language* (pp. 123-139). De Gruyter.

Srinivasan, M., & Rabagliati, H. (2015). How concepts and conventions structure the lexicon: Cross-linguistic evidence from polysemy. *Lingua*, 157, 124-152.

Entrenchment, conventionalization and cumulative culture: A usage-based perspective on language evolution

Michael Pleyer[*1], Jakob Neels[2], and Stefan Hartmann[3]

[*]Corresponding Author: pleyer@umk.pl
[1]Center for Language Evolution Studies, Nicolaus Copernicus University in Toruń, Toruń, Poland
[2]Institute of British Studies, Leipzig University, Leipzig, Germany
[3]German Department, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

This paper discusses how concepts from usage-based linguistics can prove fruitful in investigating the evolution of language. In particular, we outline recent developments in usage-based approaches to language and explore how they can inform an account of how fully-fledged language emerged from protolinguistic communication. Specifically, we focus on the concepts of entrenchment and conventionalization as well as their interaction in processes of language change and grammaticalization, and we discuss whether and to what extent such concepts can also account for the emergence of structure in hominin interactions.

1. Introduction

Recent years have seen increasing parallels in the theoretical developments within evolutionary linguistics and usage-based linguistics (e.g. Verhagen, 2021; Pleyer & Hartmann, 2024). In particular, both have converged on the view that the development of fully-fledged human language can be conceived of as the cultural evolution of a complex adaptive system, i.e., a system whose global characteristics emerge from myriads of independent interactions at a more local level (e.g. Beckner et al., 2009; Kirby, 2012; Steels, 2011). In addition, both approaches have increasingly stressed the importance of interaction as the "core ecology for language use" (Levinson & Holler, 2014), which must have played a significant role in how language emerged (e.g. Scott-Phillips, 2015). Given these parallel developments, here we explore in more detail the implications of a usage-based perspective for language evolution, focusing on theoretical frameworks that have been proposed in usage-based linguistics fairly recently.

Usage-based theory sees language structure as arising from interactional and cognitive factors operating on repeated instances of actual language usage.

Linguistic systems are conceived of as dynamic networks of symbolic form–meaning pairings, i.e., constructions, and they are recognised to be shaped by the needs and biases of communication, social interaction, cultural transmission and human cognition (Croft, 2000; Diessel, 2019; Kirby, Tamariz, Cornish & Smith, 2015). Usage-based approaches have proven successful in modeling how children "construct a language" through generalizations and schematizations over repeated instances of language use (e.g. Tomasello, 2003), and how usage-based forces shape language in diachronic, historical language change (e.g. Bybee, 2010).

However, from both the perspective of usage-based approaches and evolutionary linguistics, there is no sharp distinction between the initial emergence of language, on the one hand, and the cultural evolution of language(s) on the other. This means that "there is every reason to suppose that the very first grammatical constructions emerged in the same way as those observed in more recent history" (Bybee, 2010: 202). What follows from this is that the mechanisms uncovered in usage-based approaches regarding the dynamics and mechanisms of language change can also be applied to explanations of the evolutionary emergence of language.

Specifically, we argue that usage-based mechanisms documented in language change can help explain the gradual transition on a protolinguistic–linguistic continuum after the emergence of the first (proto)constructions in hominins (see also Hartmann & Pleyer 2021). Once recurrent solutions to communicative problems in hominin interactions started to re-occur more frequently, this led to increasing degrees of entrenchment of these communicative solutions on the cognitive side, and on the social side to their diffusion and spread throughout hominin communities of practice. Shaped by usage-based factors, they were then subject to processes of cumulative cultural evolution, leading to the emergence of modern linguistic constructions that are cognitively entrenched and socially conventionalized. In addition, we argue that usage-based forces not only have the potential to explain the gradual transition towards fully-fledged languages. They also have the potential to explain how the first protolinguistic constructions emerged in interaction, thereby kickstarting the process of the cumulative cultural evolution of language (Pleyer 2023). In the remainder of this paper, we will first describe usage-based mechanisms in linguistic and cultural evolution. In particular, we will focus on the role of the processes of entrenchment and conventionalization, as highlighted in Schmid's (2020) Entrenchment-and-Conventionalization Model. Following this, we will then outline how many of the same processes can be used to explain the emergence of fully-fledged modern languages.

2. The role of entrenchment and conventionalization in linguistic evolution

Schmid's (2020) Entrenchment-and-Conventionalization Model is a recent example of a big-picture attempt to capture the effects of usage, culture and cognition on language structure. The design of Schmid's model takes the following two principal dimensions as its starting point: the cognitive level of the individual speaker on the one hand, which Schmid labels as the macro-process of *Entrenchment*, and the socio-pragmatic dimension of speech communities on the other hand, labeled *Conventionalization*:

> *Conventionalization* is the continual process of establishing and readapting regularities of communicative behaviour among the members of a speech community, which is achieved by repeated usage activities in usage events and subject to the exigencies of the entrenchment processes taking place in the minds of speakers.
>
> *Entrenchment* is the continual reorganization of linguistic knowledge in the minds of speakers, which is driven by repeated usage activities in usage events and subject to the exigencies of the conventionalization processes taking place in speech communities. (Schmid, 2020: 2)

The two dimensions meet and interact in usage, allowing for the updating and alignment of mental representations and linguistic norms. Under each of the two principal dimensions or macro-processes, a range of more atomic forces or subprocesses can be subsumed. On the cognitive side, such processes include analogy, chunking, conceptual metaphor and coding efficiency, among others. On the socio-pragmatic side, there are motivations such as social fitness and extravagance and mechanisms such as pragmatic inferencing, accommodation and diffusion.[1] The basic idea of the theoretical blueprint just outlined is that interactions and feedback loops within and between the two dimensions, as well as their subprocesses, lead to linguistic structuration.

In fact, several usage-based models (e.g. Bybee, 2010; Traugott & Trousdale, 2013; Schmid, 2020) converge on the idea that multiple cognitive and socio-pragmatic forces, including those listed above, are at work to advance structuration and diachronic grammaticalization. In the minds of individual speakers, recurrent linguistic sequences fuel chunking. In chunked representations, syntagmatic associations are strengthened while paradigmatic associations to other instances of the chunk's lexical components are weakened. This allows emergent grammatical constructions to emancipate from their concrete lexical sources and take on a life cycle of their own. Over time, grammaticalizing constructions acquire more abstract, schematic meanings through pervasive thought processes such as metaphor and metonymy (e.g. Heine

[1] The conceptual and terminological choices that Schmid makes to characterize these forces are specific to his particular model, but they are in line with mechanisms and motivations that are widely recognized in the usage-based literature and which we decided to adopt in the present paper.

et al., 1991). In the textbook example of the English auxiliary *gonna*, the abstract target meaning of futurity is related to the source meaning of motion not only through the fundamental TIME-IS-SPACE metaphor but also through the metonymic link between moving with an intention (‘be going in order to’) and likely future events. Conceptual metonymy (i.e., accessing a target concept via another salient, experientially closely related concept) is, moreover, one of the forces that links the cognitive dimension of change and the socio-pragmatic dimension, as many steps of grammaticalization depend on speaker–hearer interaction. Metonymy structures conceptualization while hearers are inclined to draw rich pragmatic inferences as part of efficient, cooperative communication (Panther & Thornburg, 2003; Traugott, 1988). Thereby, grammaticalizing constructions can assume new procedural functions as interactants negotiate meanings in context. In socially situated accommodation, communication partners often converge on similar structures, which allows novel structures to become more than one-off patterns in a single speaker and to be replicated by others in future usage events (e.g. Brône & Zima, 2014) When proving structurally and/or socially effective, linguistic innovations diffuse to more contexts and users. They usually begin to diffuse in local social networks and tight communities of practice before being propagated into wider communities (Milroy, 1980; Nevalainen et al., 2011). Increasingly frequent use leads to “inflationary” effects (Dahl, 2001) whereby grammaticalizing constructions lose in pragmatic and semantic value. As their meaning contributions become discursively secondary and their syntagmatic predictability increases, grammaticalizing constructions also tend to reduce in phonetic substance. This results in the typical cline of increasing morphological bondedness, with nouns and verbs transforming into unstressed function words and ultimately into inflectional affixes. Overall, usage-based theoretical approaches such as Schmid’s (2020) Entrenchment-and-Conventionalization Model demonstrate that dynamic linguistic systems can be explained comprehensively based chiefly on humans’ general cognitive capacities and general socio-pragmatic/cultural processes. The channel of interaction for the individual mind and communal norms is the socially situated use and exchange of structures in repeated usage events.

3. The dynamics of cumulative culture and usage-based forces in language evolution

The cognitive, interactional and social forces attested in diachrony as a form of cumulative cultural evolution can help explain the gradual transition of protolanguage to modern language. However, a usage-based approach to language evolution can go further than that by also identifying a central locus of the emergence of linguistic structure: that of usage and interaction. That is, one crucial starting point for the emergence, diffusion and eventual conventionalization of structured communicative patterns is that they emerge as successful communicative strategies within an interaction.

In a complex adaptive system view, we can locate these processes on different interacting and connected timescales (cf. e.g. Enfield, 2014; Kirby, 2012; Steels, 2011). On the diachronic or "glossogenetic" (Hurford, 1990) timescale of cultural-historical change, new linguistic constructions emerge and become conventionalized within a population. The timescale that fuels and feeds into these cultural-historical processes is the "enchronic" (Enfield, 2014), or interactional timescale. On this timescale, new construction patterns emerge through social interactive and cognitive processes over the timespan of a conversation. It therefore represents a puzzle piece that links processes of grammaticalization and cumulative cultural evolution with the process that creates the "reusable material" for these processes.

4. Entrenchment, conventionalization and cumulative culture in language evolution

On the view presented here, usage-based forces of entrenchment and conventionalization lead to the cumulative cultural evolution of language. The likelihood of the re-emergence of particular structures would be boosted by their usefulness in interaction, leading to their increasing consolidation and entrenchment in memory. This would have also made them more likely to be used with different communicative partners, which in turn would lead to these structures emerging and spreading throughout the community. Such structures would gain the status of tacit norms: expected ways of jointly solving particular communicative challenges, a process Schmid (2020) refers to as "usualization." With increasing usualization and the repetition of particular usage activities in interactive encounters, emergent patterns would increasingly become conventionalized. They would therefore represent "what has been ritualized from interactions" (Thompson & Cooper-Kuhlen, 2005). This view is supported by simulations of interacting agents (Barr, 2004). These agents were shown to establish and maintain a shared symbolic conventional system by updating their behavior based on local, dyadic interactions instead of by adhering to system-level, global information. This view is also supported by research showing that systematic structure can emerge in communities of interactants over multiple repeated encounters without the need for generational transmission and turnover (e.g. Fay et al., 2010; Nölle et al., 2018; Raviv et al., 2019).

On the cognitive side, processes of entrenchment could also have introduced one of the key features of the complex network of constructions characterizing modern languages: the fact that these networks contain form–meaning pairings of different degrees of schematicity/abstractness and complexity (e.g. Goldberg, 2003; Stefanowitsch & Flach, 2017). Constructions range from fully concrete, specific constructions such as word constructions (*Australia*, *armadillo*), to more abstract and complex constructions such as affix schemas (such as [STEM]-[AFFIX] as in *bloody*, *colorful*) and the ditransitive construction (SUBJ V OBJ1 OBJ2, such as *I put a shrimp on the barbecue*). In the process of becoming more habitual, automated and entrenched, complex constructions often become more schematic. The usage profile of a construction expands incrementally, influencing its mental representation to become more productive and general (De Smet, 2016; Neels, 2020). This process could also explain how emergent protoconstructions in hominin communities of practice became increasingly schematic and complex. Importantly, research in the paradigm of experimental semiotics (e.g. Galantucci, Garrod & Roberts, 2012; Nölle & Galantucci, 2023) demonstrates that the evolutionary trajectory towards complex symbolic systems can be set in motion even when interactants confronted with a communicative task start out with no shared icons or symbols at all.

5. Conclusion

In this paper, we have presented a usage-based perspective on language evolution. We outlined usage-based mechanisms in linguistic evolution, particularly as they pertain to diachronic change, especially grammaticalization. These processes operate on the cognitive level, on the one hand, and on the community level on the other, through the channel of usage. On the cognitive level, this includes the mechanisms involved in entrenchment. On the cognitive level, it includes the mechanisms involved in conventionalization.

In the context of cumulative cultural evolution, this cascade of interlocking processes can help explain how the first protolinguistic structures emerged, how they subsequently increased in complexity and structure, and how they spread through communities. They therefore have the potential to explain two key aspects of language evolution from a usage-based perspective: a) the first emergence of (proto)linguistic structures; b) the gradual transition toward the modern human language pole on the protolinguistic–linguistic continuum through cumulative changes in hominin communities over a long period of time. Temporary, emergent communicative routines turned into inventories of firmly entrenched and community-wide communicative routines: protolanguages. These communication systems developed increasing degrees of conventionalization and accumulated innovations and wider contexts of use through processes of cumulative cultural change, evolving into fully grammaticalized and conventionalized structured inventories of constructions shared by communities of practice: languages.

Acknowledgements

This research is part of the project No. 2021/43/P/HS2/02729 co-funded by the National Science Centre and the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 945339. For the purpose of Open Access, the author has applied a CC-BY public copyright licence to any Author Accepted Manuscript (AAM) version arising from this submission.

References

Barr, D. J. (2004). Establishing conventional communication systems: Is common knowledge necessary? *Cognitive Science*, *28*(6), 937-962.

Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., Holland, J., Ke, J., Larsen-Freeman, D., & Schoenemann, T. (2009). Language is a Complex Adaptive System: Position Paper. *Language Learning*, *59* Suppl. 1. https://doi.org/10.1111/j.1467-9922.2009.00533.x

Brône, G., & Zima, E. (2014). Towards a dialogic construction grammar: Ad hoc routines and resonance activation. *Cognitive Linguistics*, *25*(3), 457–495. https://doi.org/10.1515/cog-2014-0027

Bybee, J. (2010). *Language, usage and cognition*. Cambridge University Press. https://doi.org/10.1017/CBO9780511750526

Croft, W. (2000). *Explaining language change: An evolutionary approach*. Pearson Longman.

Dahl, Ö. (2001). Inflationary effects in language and elsewhere. W Frequency and the Emergence of Linguistic Structure (pp. 471-480). John Benjamins Publishing Company. https://doi.org/10.1075/tsl.45.24dah

De Smet, H., & Fischer, O. (2017). The role of analogy in language change: Supporting constructions. In M. Hundt, S. Mollin, & S. E. Pfenninger (Eds.), *The changing English language: Psycholinguistic perspectives* (pp. 240–268). Cambridge University Press. https://doi.org/10.1017/9781316091746.011

Diessel, H. (2019). *The grammar network: How linguistic structure is shaped by language use*. Cambridge University Press.

Enfield, N. J. (2014). *Natural Causes of Language: Frames, biases and cultural transmission*. Language Science Press.

Fay, N., Arbib, M., & Garrod, S. (2013). How to Bootstrap a Human Communication System. *Cognitive Science*, *37*(7), 1356–1367. https://doi.org/10.1111/cogs.12048

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

Goldberg, A. E. (2003). Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences*, 7(5), 219-224. https://doi.org/10.1016/S1364-6613(03)00080-9

Hartmann, S., & Pleyer, M. (2021). Constructing a Protolanguage: Reconstructing Prehistoric Languages in a Usage-Based Construction Grammar Framework. P*hilosophical Transactions of the Royal Society B: Biological Sciences*, 376, 20200200. https://doi.org/10.1098/rstb.2020.0200

Heine, B., Claudi, U., & Hünnemeyer Friederike. (1991). *Grammaticalization: A conceptual framework*. University of Chicago Press.

Hurford, J. R. (1990). Nativist and functional explanations in language acquisition. In I. M. Roca (Ed.), *Logical issues in language acquisition* (pp. 85-136). Foris.

Kirby, S. (2012). Language is an Adaptive System: The Role of Cultural Evolution in the Origins of Structure. In M. Tallerman & K. R. Gibson (Eds.), *The Oxford Handbook of Language Evolution*. Oxford University Press.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition, 141*, 87–102.

Levinson, S. C., & 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/http://dx.doi.org/10.1098/rstb.2013.0302

Milroy, L. (1980). *Language and social networks*. Blackwell.

Neels, J. (2020). Lifespan change in grammaticalisation as frequency-sensitive automation: William Faulkner and the *let alone* construction. *Cognitive Linguistics*, *31*(2), 339–36-5. https://doi.org/10.1515/cog-2019-0020

Nevalainen, T., Raumolin-Brunberg, H., & Mannila, H. (2011). The diffusion of language change in real time: Progressive and conservative individuals and the time depth of change. *Language Variation and Change*, *23*(1), 1–43. https://doi.org/10.1017/S0954394510000207

Nölle, J., & Galantucci, B. (2023). Experimental Semiotics: Past, present, and future. In A. M. G. A. Ibáñez (Ed.), *The Routledge Handbook of Semiosis and the Brain* (pp. 66-81). Routledge.

Nölle, J., Staib, M., Fusaroli, R., & 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

Panther, K.-U., & Thornburg, L. L. (2003). *Metonymy and pragmatic inferencing*. John Benjamins Publishing Company. https://doi.org/10.1075/pbns.113

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. & Hartmann, S. (2024). *Cognitive linguistics and language evolution* (Cambridge Elements in Cognitive Linguistics). Cambridge University Press. https://doi.org/10.1017/9781009385022

Raviv, L., Meyer, A., & 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

Schmid, H.-J. (2020). *The dynamics of the linguistic system: Usage, conventionalization, and entrenchment*. Oxford University Press.

Scott-Phillips, T. C. (2015). *Speaking our Minds: Why Human Communication is Different, and How Language Evolved to Make it Special*. Palgrave Macmillan.

Steels, L. (2011). Modeling the cultural evolution of language. *Physics of Life Review*, 8, 339-356.

Stefanowitsch, A., & Flach, S. (2017). A corpus-based perspective on entrenchment. In H.-J. r. Schmid (Ed.), *Entrenchment and the psychology of language learning: How we reorganize and adapt linguistic knowledge* (pp. 101-127). De Gruyter.

Thompson, S. A., & Couper-Kuhlen, E. (2005). The clause as a locus of grammar and interaction. *Discourse Studies*, *7*(4-5), 481-505. https://doi.org/10.1177/1461445605054403

Tomasello, M. (2003). Constructing a language: a usage-based theory of language acquisition. Harvard University Press.

Traugott, E. C. (1988). Pragmatic strengthening and grammaticalization. In S. Axmaker, A. Jaisser, & H. Singmaster (Eds.), *Proceedings of the fourteenth annual meeting of the Berkeley Linguistics Society* (pp. 406–416). BLS.

Traugott, E. C., & Trousdale, G. (2013). *Constructionalization and constructional changes*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199679898.001.0001

Verhagen, A. (2021). *Ten lectures on cognitive evolutionary linguistics* (Distinguished Lectures in Cognitive Linguistics 24). Brill.

Neural repurposing as a driving force in the Baldwinian coevolution of emotional and propositional communications

Piotr Podlipniak*[1]

*Corresponding Author: podlip@amu.edu.pl
[1]Department of Musicology, Adam Mickiewicz University, Poznań, Poland

Natural speech is composed of two substantially different tools to transmit meaning i.e. affective prosody and the phonological system (Ackermann et al., 2014). Affective prosody mainly uses sounds to elicit a particular emotional state in the recipient, whilst the function of the phonological system is to transfer sound patterns into concepts – propositions. These two forms of communication are inseparable parts of every known natural speech. As speech prosody most likely has a genetic basis (Scartozzi et al., 2023) and shares many features with music, their ontogenetic development starts from a common phase (McMullen & Saffran, 2004), therefore many researchers have proposed that natural language evolved from a music-like protolanguage (Bannan, 2008; Brown, 2000, 2017; Fitch, 2005, 2013, Mithen, 2006).

However, although affective prosody is evolutionarily older than the phonological system and is shared by a broad group of mammalian species (Filippi, 2016), prosodic features such as pitch contour, stress, and timing are also more or less involved in the transmission of propositional meaning in speech (Nygaard et al., 2009). Yet, such an involvement is hardly present in music, probably because music is devoid of combinatorial symbolic meaning. Apart from this, natural language can take the form of sign language in which all speech properties are implemented (transduced) into the gestural domain. It has also been observed that non-human primates exchange propositional meaning by the means of culturally invented gestures (Hobaiter & Byrne, 2014) and vocalizations (Wright et al., 1990), which suggests that hominins also used some kinds of propositional communication independent of affective prosody. These observations challenge the hypotheses of a single common ancestor of language and music, and their subsequent linear independent evolution.

In this presentation, an alternative view is proposed in which affective prosody, sound symbols, and pantomime were all parts of a hominin communicative niche, but none of them were the main precursor of language. Instead, by being initially independent and different in terms of their communicative mechanisms, such as directly eliciting emotions (affective prosody), symbolizing sounds, and pantomime gesturing, they started to interact in response to new selective pressures resulting from increasing social complexity. On the one hand, unconscious and direct induction of emotions could have been involved in fulfilling adaptive functions such as strengthening social bonds (Savage et al., 2021) or in free rider recognition (Podlipniak, 2023). On the other hand, symbolic, indexical, and iconic tools could have developed in response to the need of selecting, amplifying, and sharing specific thoughts as well as planning the future. These functions could have led to the coevolution of consciousness (Dehaene, 2014) and natural language. As a driving force in this process, a proximal mechanism of 'neural repurposing' is proposed. Neural repurposing is a specific kind of exaptation (Gould & Vrba, 1982) and consists of reusing existing neural circuitry in a functionally novel neural tool (Schlaudt, 2022). Such a change can be culturally induced and achieved by neural plasticity which means that cultural invention could have been a source of interactions between affective prosody, sound symbols and pantomime among hominins. In fact, neural repurposing has been discovered in the contemporary communicative domain. For instance, it has been observed that native speakers of tonal languages differ in the lateralization of pitch processing from non-native speakers (Gu et al., 2013; Li et al., 2021; Liang & Du, 2018). The change of lateralization has also been noticed among users of Turkish whistle language (Güntürkün et al., 2015) – a culturally invented form of distance communication. Taking into account that pitch contour is a widely used clue to indicate grammatical mood (Jun, 2005; Warren & Calhoun, 2021), and prosodic accents are also an effective tool to communicate the hierarchy of words, it seems to be reasonable to assume that the elements of emotional communication were repurposed in order to fulfil new functions in the exchange of propositional meaning. Similarly, the use of pantomime to transmit propositional meaning opened the way to the emergence of conventionalized gestures (Zlatev et al., 2020). However, due to biological costs burdened on strenuous learning of these new hominin expressions, natural selection started to favor individuals (and their progeny) that were accidentally endowed with the predisposition to learn these expressions faster and less strenuously. This process – Baldwinian evolution (Baldwin, 1896a, 1896b) – could have led to the genetic canalization of the use of prosody in the transmission of propositional meaning.

References

Ackermann, H., Hage, S. R., & Ziegler, W. (2014). Brain mechanisms of acoustic communication in humans and nonhuman primates: An evolutionary perspective. *Behavioral and Brain Sciences*, 37(06), 529–546.

Baldwin, J. M. (1896a). A New Factor in Evolution. *The American Naturalist*, 30(354), 441–451.

Baldwin, J. M. (1896b). A New Factor in Evolution (Continued). *The American Naturalist*, 30(355), 536–553.

Bannan, N. (2008). Language out of Music: The Four Dimensions of Vocal Learning. *The Australian Journal of Anthropology*, 19(3), 272–293.

Brown, S. (2000). The ‘musilanguage’ model of musical evolution. In N. L. Wallin, B. Merker, & S. Brown (Eds.), *The origins of music* (pp. 271–300). The MIT Press.

Brown, S. (2017). A Joint Prosodic Origin of Language and Music. *Frontiers in Psychology*, 8, 1894.

Dehaene, S. (2005). Evolution of Human Cortical Circuits for Reading and Arithmetic: The ‘Neuronal Recycling’ Hypothesis. In S. Dehaene, J.-R. Duhamel, M. D. Hauser, & G. Rizzolatti (Eds.), *From Monkey Brain to Human Brain: A Fyssen Foundation Symposium*. The MIT Press.

Dehaene, S. (2014). *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts*. Penguin books Ltd.

Filippi, P. (2016). Emotional and interactional prosody across animal communication systems: A comparative approach to the emergence of language. *Frontiers in Psychology*, 7, 1393.

Fitch, W. T. (2005). Protomusic and protolanguage as alternatives to protosign. *Behavioral and Brain Sciences*, 28(2), 132–133.

Fitch, W. T. (2013). Musical protolanguage: Darwin’s theory of language evolution revisited. In J. J. Bolhuis & M. Everaert (Eds.), *Birdsong, speech, and language: exploring the evolution of mind and brain* (pp. 489–503). The MIT Press.

Gould, S. J., & Vrba, E. S. (1982). Exaptation-A Missing Term in the Science of Form. *Paleobiology*, 8(1), 4–15.

Gu, F., Zhang, C., Hu, A., & Zhao, G. (2013). Left hemisphere lateralization for lexical and acoustic pitch processing in Cantonese speakers as revealed by mismatch negativity. *NeuroImage*, 83, 637–645.

Güntürkün, O., Güntürkün, M., & Hahn, C. (2015). Whistled Turkish alters language asymmetries. *Current Biology*, 25(16), R706–R708.

Hobaiter, C., & Byrne, R. W. (2014). The Meanings of Chimpanzee Gestures. *Current Biology*, 24(14), 1596–1600.

Jun, S.-A. (2005). Prosodic Typology. In S.-A. Jun & S.-A. Jun (Eds.), *Prosodic Typology: The Phonology of Intonation and Phrasing* (pp. 430–458). Oxford University Press.

Li, Y., Tang, C., Lu, J., Wu, J., & Chang, E. F. (2021). Human cortical encoding of pitch in tonal and non-tonal languages. *Nature Communications*, 12(1), 1161.

Liang, B., & Du, Y. (2018). The Functional Neuroanatomy of Lexical Tone Perception: An Activation Likelihood Estimation Meta-Analysis. *Frontiers in Neuroscience*, 12, 495.

McMullen, E., & Saffran, J. R. (2004). Music and Language: A Developmental Comparison. *Music Perception*, 21(3), 289–311.

Mithen, S. J. (2006). *The singing Neanderthals: the origins of music, language, mind, and body*. Harvard University Press.

Nygaard, L. C., Herold, D. S., & Namy, L. L. (2009). The Semantics of Prosody: Acoustic and Perceptual Evidence of Prosodic Correlates to Word Meaning. *Cognitive Science*, 33(1), 127–146.

Podlipniak, P. (2023). Free rider recognition—A missing link in the Baldwinian model of music evolution. *Psychology of Music*, 51(4), 1397– 1413.

Savage, P. E., Loui, P., Tarr, B., Schachner, A., Glowacki, L., Mithen, S., & Fitch, W. T. (2021). Music as a coevolved system for social bonding. *Behavioral and Brain Sciences*, 44, e59: 1-22.

Scartozzi, A., Gustavson, D., Creanza, N., Magne, C., Below, J., Gordon, R., & Nayak, S. (2023). The genetics of speech prosody perception: genetic associations with communication-related traits - poster at World Congress of Psychiatric Genetics October 10-14, 2023. *European Neuropsychopharmacology*, 75, S170.

Schlaudt, O. (2022). Exaptation in the Co-evolution of Technology and Mind: New Perspectives from Some Old Literature. *Philosophy & Technology*, 35(2), 48.

Warren, P., & Calhoun, S. (2021). Intonation. In J. Setter & R.-A. Knight (Eds.), *The Cambridge Handbook of Phonetics* (pp. 209–236). Cambridge University Press.

Wright, A. A., Shyan, M. R., & Jitsumori, M. (1990). Auditory same/different concept learning by monkeys. *Animal Learning & Behavior*, 18(3), 287–294.

Zlatev, J., Żywiczyński, P., & Wacewicz, S. (2020). Pantomime as the original human-specific communicative system. *Journal of Language Evolution*, 5(2), 156–174.

Movement-related muscle activity and kinetics affect human vocalization amplitude

Wim Pouw*[1], Lara Burchardt [1, 2], Raphael Werner[1], and Luc Selen[1]

*Corresponding Author: wim.pouw@donders.ru.nl
[1]Donders Institute for Brain, Cognition and Behavior, Radboud University Nijmegen
[2]Leibniz Institute for General Linguistics ZAS Berlin

Voice production can be a whole-body affair: Upper limb movements physically impact the voice in steady-state vocalization, speaking, and singing. This is supposedly due to biomechanical impulses on the chest-wall, affecting subglottal pressure. Unveiling such biomechanics is important, as humans gesture with their hands in a synchronized way with speaking. Here we assess biomechanical interactions between arm movements and the voice, by measuring activity of respiratory-related muscles during different types of upper limb movement. We show that positive peaks in the voice's amplitude increase with movements and more so with a 1 kg weight attached to the wrist. We further report exploratory findings that gesture-related muscle activations scale with positive peaks in the voice's amplitude. These results indicate that the voice aligns with the forces generated by the body and implies that the voice evolved in the context of bodily action.

1. Introduction

In principle, any muscle attached to the human rib cage can act on it and thus affect the subglottal pressure that supports voicing. Consequently, there are many potential respiratory-vocal muscles. Said muscles would include those around the chest (e.g., pectoralis major), abdomen (rectus abdominis), and back (erector spinae, serratus posterior/anterior). Passive breathing and speaking, however, is mainly driven by the diaphragm and the intercostal muscles between the ribs (Aliverti, 2016; Seikel et al., 2019). Only on rare occasions, when coughing, shouting, or breathing deeply, humans recruit other so-called "accessory" respiratory muscles such as the abs and pectoral muscles (Aliverti, 2016; Lasserson et al., 2006; Seikel et al., 2019).

Yet, when humans speak or sing, it is common to move the upper limbs expressively at the same time – called gesture (Pearson & Pouw, 2022; Wagner et al., 2014). Such upper limb movements will recruit a whole range of upper body muscles, including those involved in maintaining posture (Cordo & Nashner, 1982). Several of these muscles attach to the rib cage (e.g., abdominal and pectoral muscles) and are classically listed as accessory to respiratory functioning (Seikel et al., 2019). Given that speaking requires subtle modulations of subglottal pressure (Rubin et al., 1967; Sundberg et al., 1993a; Sundberg et al. 1993b), co-gesture speaking must be in some way coordinated with the respiratory-and-thus-vocal muscles that are activated during gesturing (see Pouw & Fuchs, 2022).

There is considerable evidence that gestural arm movements affect the voice directly, as summarized by Pouw et al. (2019b): More extreme peaks in the acceleration of movements bigger (arm) and smaller (wrist) upper limb movements relate to more chest-circumference changes, which is associated with more extreme acoustic effects on the intensity of vocal sound (via increasing subglottal pressure). Furthermore, acoustic effects of upper limb movements are more pronounced when subjects are in a less stable standing vs. sitting position (Pouw et al., 2019b). This all ties in with the idea that a physical impulse (mass x acceleration), impacts posture (especially when standing), recruiting respiratory-related muscles (that change chest circumference), which impacts respiratory-vocal functioning (such that intensity is affected).[1] These previous gesture-speech physics[2] studies assessed continuous voicing, mono-syllable utterances, and fluent speech production (Pouw et al., 2020). However, direct evidence of mass and muscle activity relating to gesture-speech physics has so far not been reported.

We ask two questions here: 1) Do different upper limb movements lead to dissociable positively peaked deviations of the amplitude envelope of ongoing voicing? 2) Does peak muscle activity predict positively peaked deviations of the amplitude envelope of ongoing voicing? In addition to the results from the two pilot participants, we also report confirmatory pre-registered results with respect to the first question assessing differences in positive amplitude peaks.

Based on the gesture-speech physics account (Pouw & Fuchs, 2022) we predict that movements will increase the magnitude of positive amplitude peaks in the voice and that we should find (for some muscles) that peak muscle activity relates to the amplitude envelope.

2. Method

We report exploratory results with N = 2 participants, and some confirmatory pre-registered results (N = 17).[3] For the current pilot experiment supporting the pre-registration, the first author (Dutch-speaking; male; BMI = 21.7) and a volunteer

[1] Such interactions, between pectoral/upper limb activity and respiratory-vocal states, have been well-studied in non-human animals (Cooper & Goller, 2004; Lancaster et al., 1995; Blumberg, 1992).

[2] Gesture-speech physics (Pouw et al., 2019a) does not require the speaker to move to modulating vocalization, but conversely moving may affect vocalizations, or even voiceless expirations (Werner et al., 2024).

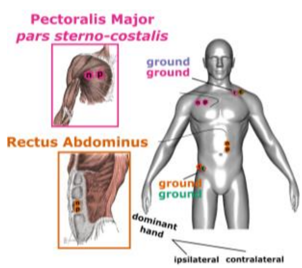
[3] The current study has been approved by the Ethics Committee Social Sciences (ECSS) of the Radboud University (reference nr.: 22N.002642).

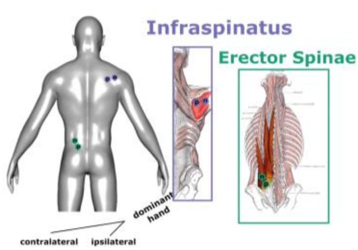
female (Dutch-speaking; BMI = 21.5) performed the experiment.[4] We also report a selection of the confirmatory results we pre-registered. This full dataset consisted of N = (17) participants (7 f, 10 m), M (SD) age = 28.5 (6.5), BMI = 23.40 (2.20).

The 1-hour study involved a two-level within-subject factor wrist-weight manipulation (no weight, 1 kg weight), a two-level within subject vocalization condition (expire, vocalize), and a five-level within-subject movement condition (‘no movement’, ‘elbow extension’, ‘elbow flexion’, ‘shoulder external rotation’, ‘shoulder internal rotation’). With 4 trial repetitions over the experiment, we yield 80 trials per participant. Trials were blocked by weight condition and vocalization condition. Within blocks, all movement conditions were randomized.

To manipulate the mass set in motion, we apply a wrist weight. We use a TurnTuri sports wrist weight of 1 kg. The experiment was coded in Python using functions from PsychoPy. The participants were recorded via a video camera. We used Mediapipe (Lugaresi et al., 2019) to track the skeleton and facial movements, which is implemented in Masked-piper which we also use for masking the videos (Owoyele et al., 2022).

We measured surface ElectroMyoGraphy (sEMG) using a wired BrainAmp ExG system (sampling rate: 2,500 Hz). Disposable surface electrodes were used, and for each of the four target muscles we had 3 (positive, negative, ground) electrodes. Positive and negative electrodes were attached with a 15 mm distance center to center. We applied electrodes for focal muscles which directly participate in the internal (pectoralis major) and external rotation (infraspinatus) of the humerus. We attached the electrodes for focal muscles ipsilaterally (relative to the dominant hand) to the muscle belly of the clavicular head of the pectoralis major, with a ground electrode on the clavicle on the opposite side. We recorded postural muscles: applying electrodes to muscles that anticipate and react to postural perturbations due to upper limb movements. Electrodes for these muscles were attached contralaterally to the moving dominant hand to the rectus abdominis and the erector spinae muscle group (specifically, the iliocostalis romborum), with ground electrodes on the iliac crest on the opposite side.





For audio recordings, we used a headset microphone sampling at 16 kHz. The gain levels of the condenser power source were set by the hardware. We also record ground reaction forces, but these will not be discussed here.

[4] More detail on participants, equipment, and data processing can be found in the pre-registration.

We use LabStreamLayer as an interface for data synchronization across signals. After body measurements, we applied the surface EMG. After practice trials, participants performed 80 blocked trials. For each trial, participants were closely guided by the information on the monitor.

Participants were instructed to adopt the start position of the movement, which is a 90° elbow flexion, with either an externally rotated humerus (for internal rotation), or a non-rotated humerus with the wrist in front of the body (rest position for the other movement conditions). For the no movement condition participants were asked to rest their arms alongside their bodies. Upon trial start, participants inhaled deeply with a timer counting down from 4 seconds. Subsequently, participants were asked to continuously 'vocalize' with a *schwa* sound, or 'expire', with a screen appearing after 3 seconds to perform the movement with visual guidance to where the movement's end position is so that participants are reminded of the movement. After an additional 4 seconds, the trial ends, which gives enough time to perform the movement and stabilize vocalization after the perturbation. Participants were explicitly instructed to keep their vocalization as stable as possible during the different movement conditions. To cater for vocal amplitude decrease as the lungs deflate over time, we detrended the amplitude envelope time series and expressed positive peaks relative to this trend line.

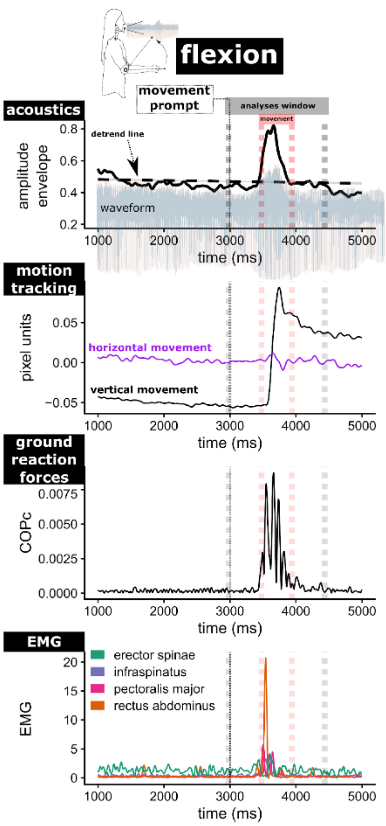


Figure 2: Time series example for a flexion movement + vocalization trial.

We only analyze vocalization trials here.

An example time series is shown in Fig. 2. At time = 0, the prompt is given to the participant to vocalize. We determine a detrending line using linear regression for the 1 to 5 s after the vocalization prompt. At 3,000 ms there is a movement prompt. We assess signal peaks in a time window running from 500 ms before movement onset to 500 ms after movement offset, as indicated by gray dashed bars.

3. Exploratory Results

3.1 Effects of different movements on positive peaks in vocalization

We first modeled with a mixed linear regression the variation in positive peaks in the amplitude envelope (using `R`-package `lme4`), with participant as random intercept.[5] The model coefficients are given in Table 1. There is a positive, but not statistically reliable effect of wrist weight in this *exploratory* sample. Further, all movements (extension, flexion, internal rotation, external rotation) lead to statistically reliable increases in positive peaks in the vocalization amplitude envelope relative to the no movement condition (with flexion and external rotation leading to more extreme effects).

Table 1: Effects of weight and movement condition on positive peaks in the voicing amplitude envelope.

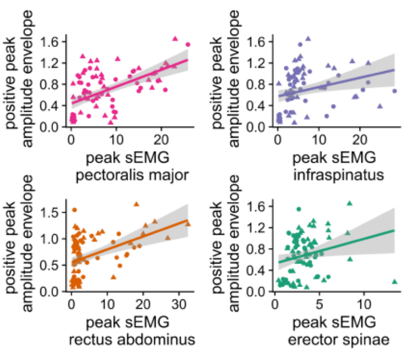
	Est.	SE	t-value	p-value
Intercept	0.164	0.101	1.63	0.181
vs. weight	0.062	0.063	0.98	0.328
vs. extension	0.436	0.100	4.35	<.001
vs. flexion	0.618	0.100	6.17	<.001
vs. internal r.	0.568	0.100	5.67	<.001
vs. external r.	0.794	0.100	7.92	<.001

3.2 Effects of muscle activity on positive peaks in vocalization

We also directly relate muscle activity peaks with the positive peaks in the amplitude envelope (after checking for collinearity). The model coefficients (for a model with the different peak muscle activities as predictor and participant as random intercept), are given in Table 2 and show that peak EMG activity in all the muscles (but especially the rectus abdominis, a well-known expiratory muscle) leads to statistically reliable increases in positive peaks in the amplitude envelope.

Table 2 and Figure 3: Muscle activity effects on magnitude positive peaks in vocalization. In the figure triangles indicate movements with wrist-weight.

	Est.	SE	t-value	p-value
Intercept	0.311	0.094	3.315	0.072
erector spinae	-0.019	0.017	-1.112	0.270
infraspinatus	0.009	0.005	1.825	0.072
pectoralis major	0.029	0.005	5.973	<.001
rectus abdominis	0.030	0.006	5.266	<.001



4. Confirmatory Results Research Question 1

We here report the *confirmatory* results for the effects of different movements on positive peaks in vocalization, in a model including weight and movement

---

[5] Every model reported was compared to a base model predicting the overall mean and explained more variance than the base model.

conditions. We confirm that different types of movement, as well as adding a weight to the moving wrist, increases the positive peaks in the amplitude envelope relative to a no movement/no weight condition.

	Est.	SE	t-value	p-value
Intercept	0.233	0.037	6.361	<.001
vs. weight	0.058	0.021	2.825	0.005
vs. extension	0.130	0.032	4.005	<.001
vs. flexion	0.164	0.033	5.057	<.001
vs. internal r.	0.104	0.033	3.204	0.001
vs. external r.	0.116	0.032	3.586	<0.001

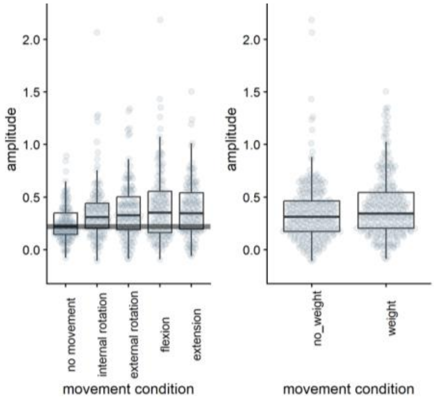


Table 4 & Figure 5: Confirmatory effects of weight and movement condition on positive peaks in the amplitude envelope.

5. Discussion & Conclusion

In summary: Movement of the upper limb yields unintentional positive peaks on the amplitude envelope of vocalization. This can be labeled 'unintentional' as the task is to produce a stable vocalization output. Further, we show promising exploratory results that activity of specific muscles is reliably related to these positive peaks in the voice’s amplitude. Though in the exploratory sample we did not find an effect of weight, in the confirmatory study we were able to confirm this small but reliable effect of wrist weight on voice peaks, suggesting that the vocalization peaks are related to the required forces (kinetics) to move the segment and not necessarily to the movement itself (i.e., to the muscle activity involved; Pouw et al., 2019b).

The results reported here will be further confirmed by analyzing the data regarding muscular and postural effects on the voice (Pouw et al., 2023). Such analyses will illustrate whether it is possible to make clear predictions about what type of upper limb movements have a certain effect on the voice that could then be functionally integrated with speech. To conclude, we hereby show that voice production is a dynamically open system that will be affected by other communicative actions such as hand gestures. This has deep implications for why gesture and speech are often produced in synchrony in humans in specific, and how evolution of the mammalian voice in general might be related to the whole-body movement system (Pouw & Fuchs, 2022). Voices do not operate in a vacuum, they are produced by bodily elements that can be synergistically recruited for moving one’s body too.

References

Aliverti, A. (2016). The respiratory muscles during exercise. *Breathe*, *12*(2), 165-168.

Blumberg, M. S. (1992). Rodent ultrasonic short calls: locomotion, biomechanics, and communication. *Journal of Comparative Psychology*, *106*(4), 360-365.

Cooper, B. G., & Goller, F. (2004). Multimodal signals: enhancement and constraint of song motor patterns by visual display. *Science*, *303*(5657), 544-546.

Cordo, P. J., & Nashner, L. M. (1982). Properties of postural adjustments associated with rapid arm movements. *Journal of neurophysiology*, *47*(2), 287-302.

Lancaster, W. C., Henson Jr, O. W., & Keating, A. W. (1995). Respiratory muscle activity in relation to vocalization in flying bats. *Journal of Experimental Biology*, *198*(1), 175-191.

Lasserson, D., Mills, K., Arunachalam, R., Polkey, M., Moxham, J., & Kalra, L. (2006). Differences in motor activation of voluntary and reflex cough in humans. *Thorax*, *61*(8), 699-705.

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M. G., Lee, J., Chang, W.-T., Hua, W. Georg, M. & Grundmann, M. (2019). Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.

Owoyele, B., Trujillo, J., De Melo, G., & Pouw, W. (2022). Masked-Piper: Masking personal identities in visual recordings while preserving multimodal information. *SoftwareX*, *20*, 101236.

Pearson, L., & Pouw, W. (2022). Gesture–vocal coupling in Karnatak music performance: A neuro–bodily distributed aesthetic entanglement. *Annals of the New York Academy of Sciences*, *1515*(1), 219-236.

Pouw, W., & Fuchs, S. (2022). Origins of vocal-entangled gesture. *Neuroscience & Biobehavioral Reviews*, *141*, 104836.

Pouw, W., Harrison, S. J., & Dixon, J. A. (2019a). Gesture–speech physics: The biomechanical basis for the emergence of gesture–speech synchrony. *Journal of Experimental Psychology: General*, *149*(2), 391.

Pouw, W., Harrison, S. J., Esteve-Gibert, N., & Dixon, J. A. (2019b). Energy flows in gesture-speech physics: exploratory findings and pre-registration of confirmatory analysis.

Pouw, W., de Jonge-Hoekstra, L., Harrison, S. J., Paxton, A., & Dixon, J. A. (2020). Gesture–speech physics in fluent speech and rhythmic upper limb movements. *Annals of the New York Academy of Sciences*, *1491*(1), 89-105.

Pouw, W., Werner, R., Burchardt, L. S., & Selen, L. (2023). The human voice aligns with whole-body kinetics. *bioRxiv*, 2023-11.

Rubin, H. J., LeCover, M., & Vennard, W. (1967). Vocal intensity, subglottic pressure and air flow relationships in singers. *Folia Phoniatrica et Logopaedica*, *19*(6), 393-413.

Seikel, J. A., Drumright, D. G., & Hudock, D. J. (2019). *Anatomy & physiology for speech, language, and hearing*. Plural Publishing, Incorporated.

Sundberg, J., Elliot, N., Gramming, P., & Nord, L. (1993a). Short-term variation of subglottal pressure for expressive purposes in singing and stage speech: a preliminary investigation. *Journal of Voice*, *7*(3), 227-234.

Sundberg, J., Titze, I., & Scherer, R. (1993b). Phonatory control in male singing: A study of the effects of subglottal pressure, fundamental frequency, and mode of phonation on the voice source. *Journal of Voice*, *7*(1), 15-29.

Wagner, P., Malisz, Z., & Kopp, S. (2014). Gesture and speech in interaction: An overview. *Speech communication*, *57*, 209-232.

Werner, R., Selen, L., & Pouw, W. (2024). Arm movements increase acoustic markers of expiratory flow. *bioRxiv*, 2024-01.

Cognitive biases explain constrained variation in noun classification

Ponrawee Prasertsom*[1], Kenny Smith[1], and Jennifer Culbertson[1]

*Corresponding Author: ponrawee.prasertsom@ed.ac.uk
[1]Centre for Language Evolution, University of Edinburgh, Edinburgh, United Kingdom

1. Background

A central issue in language evolution is how to explain the apparently constrained variation across the world's languages. Here, we focus on explaining the constrained variation of noun classification (Seifart, 2010): grammatical genders (as in Romance languages), noun classes and classifiers. Cross-linguistically, only some domains, such as animacy, commonly form conceptual bases for classifying nouns (e.g., Swahili noun classes), whereas other potentially salient domains such as colour never do (e.g., no "warm-coloured" classes; cf. Talmy, 1985). Following works linking cognitive biases to typology (e.g., Maldonado & Culbertson, 2022), we hypothesise that this results from a cognitive bias for animacy and/or against colour in grouping nouns. In our pre-registered experiments (OSF project: `https://osf.io/b6yns`), we test if 1) such a bias exists in noun class learning (Exp 1a-1b) and if 2) the bias is specific to language (Exp 2a-2c), as sometimes suggested in the literature (Cinque, 2013).

2. Experiments 1a-1b: Artificial noun class learning tasks

In Exp 1a, participants were randomly assigned one of two conditions (Colour and Animacy, N=40 each). In both, they were trained on the artificial nouns through images (Fig 1, left) and audio to criteria (scoring 13 out of 16 at test). Each noun may be animate (a frog/a lizard) or inanimate (a box/a bag), warm- (red/yellow) or cool-coloured (blue/green). Participants then learned two noun classes through determiners that vary based on the noun colour/animacy depending on the condition. The Animacy participants scored higher ($\beta_{\text{Cond.}} = 1.86$, $p = 0.023$, deviation-coded, mixed-effects logistic model), suggesting an animacy bias. The effect is reliable but not large (Prop. correct $0.92{\pm}0.03$ in Animacy vs. $0.84{\pm}0.04$ in Colour), perhaps because the simplicity of the language masked the bias.

In Exp 1b, we use an extrapolation design to determine whether a stronger bias emerges when participants are trained on an ambiguous system and must decide at test whether classification is based on colour or animacy. Here, the procedures remained largely the same, but participants (N=80) were not assigned to conditions. Crucially, during noun class training, the stimuli were compatible with both

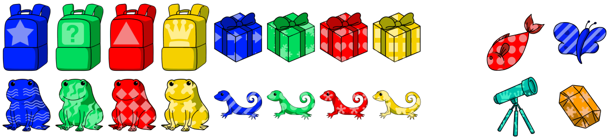


Figure 1. Visual stimuli for Exps 1a-1c, 2a, 2c (left) and for 2b (right). Only a subset is shown for 2b. stimuli to illustrate the colours and animacy types.

animacy- and colour-based systems (e.g., animates were always warm-coloured, inanimates always cool-coloured), but critical test trials asked the participants to select the determiner variant for unseen combinations (e.g., warm-coloured inanimates), forcing them to choose the criterion. The results show a strong bias to classify by animacy (Mean prop. animacy-based classification = 0.78, $p < 0.001$).

3. Experiments 2a-2c: Image sorting tasks

In Exp 2a, participants (N=30) were asked to sort the images used to represent nouns in Exps 1a-1b (Fig 1, left) into two groups. If participants have an animacy bias, they are predicted to prefer sorting by animacy to sorting by colour. This prediction was borne out, with most people (88%) sorting by animacy ($p < 0.001$, Wilcoxon signed-rank test on adjusted mutual information values).

Exps 2b-2c addressed the possibility that the stimuli were not representative of the world: Red and yellow may be unusually different for warm colours, and frogs and lizards too similar for animates. In Exp 2b, we increased the intra-category similarity in the stimuli for colour (e.g., using orange instead of yellow) and decreased it for animacy (e.g., using butterflies and fish instead of frogs and lizards). Fig 1 (right) gives a sample. In Exp 2c, we reduced the stimulus set to one category per domain (e.g., an animate is always a frog, a warm-coloured thing always red), eliminating all intra-category differences. We reproduced the animacy bias in Exp 2b (62% by animacy, 20% by colour, $p < 0.001$), but saw a different pattern in 2c (38.3% by animacy, 61.7% by colour, $p = 0.071$).

4. Discussion

Exps 1a-1b showed that an animacy bias exists in noun class learning, and could explain the prevalence of animacy and the absence of colour in noun classification. Exps 2a-2b showed that the bias is not domain-specific; it was also observed in non-linguistic categorisation. The contrast between Exp 2a-2b and Exp 2c results suggests that the animacy bias may result from the fact that an animacy-based classification offers more coherent, clear-cut clusters than a colour-based one *under variability* (Exp 2a-2b). With only one type of animacy/colour, animacy does not offer classificatory advantage (Exp 2c). When combined with the idea that cultural transmissions can amplify soft biases (Culbertson & Kirby, 2016), our results strengthen the explanatory power of cognitive biases in typology.

References

Cinque, G. (2013). Cognition, universal grammar, and typological generalizations. *Lingua*, *130*, 50–65.

Culbertson, J., & Kirby, S. (2016). Simplicity and specificity in language: Domain-general biases have domain-specific effects. *Frontiers in Psychology*, *6*.

Maldonado, M., & Culbertson, J. (2022). Person of interest: Experimental investigations into the learnability of person systems. *Linguistic Inquiry*, *53*(2), 295–336.

Seifart, F. (2010). Nominal classification. *Language and Linguistics Compass*, *4*(8), 719–736.

Talmy, L. (1985). Lexicalization patterns: Semantic structure in lexical forms. In T. Shopen (Ed.), *Language typology and syntactic description* (Vol. 3, pp. 36–149).

Survival of the *wittiest* (not *friendliest*): The art and science of human evolution

Ljiljana Progovac[*1]

[*]Corresponding Author: progovac@wayne.edu
[1]Linguistics Program, Wayne State University, Detroit, USA

Research on language evolution has largely neglected the artistic dimension of language, including eloquence and wittiness, and yet the fitness in humans has been found to be correlated with linguistic prowess, and human mate choice even today is often influenced by displays of cognitive abilities through the creatives use of language. My argument is that selection for quick-wittedness ("using *words* in a clever and funny way"), *specific to language and unique to humans*, needs to be added to the complex picture of human evolution, relevant from the earliest stages of language. Wittiness is that kind of trait which allows competition (by 'outwitting' others) while at the same time favoring "friendliness" in the sense that it provides an excellent platform for replacing physical aggression with verbal behavior. There are several previous findings, both theoretical and experimental, that have paved the way toward the view of human evolution as the "survival of the wittiest," offering better explanatory power than the "survival of the friendliest."

1. Survival of the fittest?

Much of linguistic research on language evolution has focused on human analytical abilities to form sentences with which to express (complex) propositions and thoughts. While this is certainly an important function of language(s), this is not the only function, and cannot be expected to have been a major function in the earliest stages of language. Understanding the motivation behind evolving language at these earliest stages is at the heart of understanding how language took off, and how we humans as a species evolved. While some approaches deny the role of sexual selection in the evolution of language, or the possibility of gradual, adaptive evolution (e.g. Berwick and Chomsky 2016), several others have found it advantageous to invoke adaptation and sexual selection for specifically language skills (e.g. Miller 2000; Franks and Rigby 2005; Locke 2009; Progovac and Locke 2009; Progovac 2015). Here I show that some of these latter approaches have paved the way toward a scenario favoring the "survival of the wittiest," in these earliest steps, but also beyond. But what about the survival of the fittest, or the survival of the friendliest?

Charles Darwin's work (e.g. 1859, 1872, 1874) has been associated with the "survival of the fittest," which is often (inaccurately) interpreted to be limited to the survival of the strongest, or the healthiest, physically speaking, perhaps fittest in the modern (gym) sense of the word "fitness." But "fitness" in the biological sense is a much broader concept, in that it does not single out any one specific trait. Instead, it can refer to any trait which happens to provide a better survival rate in the immediate environment, whatever that may be in that specific time or place, such as the camouflage adaptations of many species, which simply adapt to the color or shape of their environment. Having said that, adaptation in the physical sense of fitness (stronger, healthier) has certainly played an important role in the evolution of humans (and other species), and continues to do so, but see below regarding reduced physical aggression in humans.

2. **Survival of the friendliest?**

As a seeming counterproposal to supposedly ruthless physical competition, "survival of the friendliest" has recently garnered a lot of support, i.e. the view that *Homo sapiens* evolved via selection for prosociality, associated with a reduction in reactive aggression (e.g. Hare, 2017; Hare and Woods, 2020). It would appear at first sight that this is a more benevolent view of the human species, considering that it now does not seem to matter who is strong and ruthless, but who is friendly (see e.g. the *Washington Post* interview with Brian Hare (Cimons 2020), titled "'Friendliest,' not fittest, is key to evolutionary survival, scientists argue in book," implying (wrongly) that the two views are distinct (see the discussion below). However, in the scenario of the selection for "friendliness," what one might consider as "friendly" types had to have competed (ruthlessly) against "unfriendly" types. In fact, as discussed in e.g. Wrangham (2021), this would have included teaming up to kill alpha males, not a particularly friendly gesture. In the interview mentioned above, Brian Hare is quoted as saying: "We are the friendliest human species that ever evolved, which has allowed us to outcompete other human species that are now extinct … When [the mechanism] is turned on, it allows us to win. We win by cooperation and teamwork." For all our friendliness, we seem quite content with driving other species/populations to extinction, and we seem quite preoccupied with "winning" (this short quote has the word "win" in it twice).

There is no doubt that the ability to form alliances (to defeat common enemies) can be adaptive, but the term "friendliness" does not seem particularly suited for this trait. In this respect, Wrangham refers to this trait as "groupishness," characterizable as the ability to form groups and to cooperate *within* those groups, often in order to be stronger together, and better able to outcompete *other* groups or individuals. In this sense, the proposal of the survival of the friendliest (i.e. the most groupish) is not distinct from the original notion of the survival of the fittest, and it certainly does not do away with human ruthlessness. Just like physical strength, groupishness/friendliness may also be subject to sexual selection (e.g. Hare et al. 2012; Gleeson, 2018). But, crucially, and most relevant for this paper, these two aspects, selection for physical fitness, and selection for friendliness/groupishness, while both relevant for human evolution to some extent, do not begin to capture the essence of what it means to be human, specifically in relation to language. We may like physical strength in our mates, and we may like their friendly demeanor, but these are not necessarily traits that have to do with language directly, and neither are these traits unique to humans (close to home, consider e.g. bonobos, as discussed in Hare 2017). There is thus a need to supplement these general approaches to human evolution with the ones that directly and causally implicate human language.

3. **Survival of the wittiest**

My argument is that selection for quick-wittedness, specifically demonstrated by language and unique to humans, was relevant from the earliest stages of language evolution. Being witty is certainly a rather desirable trait in humans even today, a form of art we are not all equally good at, even though we all may strive to be. Wit and wittiness are characterizable as showing quick and inventive *verbal* humor, using *words* in a clever and funny way (Cambridge Dictionary; Merriam Webster Dictionary). Wittiness specifically refers to one's agility with words, including the ability to *outwit* others, i.e. to outcompete them with words. Regrettably, perhaps, being witty cannot be equated with being intelligent or wise, in the sense of making the best decisions, or solving problems in an optimal way. The latter skills are certainly adaptive, but may be harder to gauge in the context of sexual selection, as they take much longer to evaluate. Quick-wittedness is immediately there to observe and admire. It may seem like a shallow skill, like the colorful, imaginative structures built by bower birds during the mating season (Uy 2001), but it is a form of art that appeals to some deep aesthetic and emotional aspect of human existence. If we indeed come from a series of generations that sexually selected for the art (and beauty) of quick-wittedness, then we are genetically predisposed to be attracted to it, frivolous or not.

My proposal that the "survival of the wittiest" has been one of the crucial drivers of human evolution is *not* an alternative to Darwin's notion of the survival of the fittest, but rather just a specific rendering of this approach when applied specifically to language evolution in humans. The argument I defend here is that, since the emergence of language, and to this day, the fitness in humans has been highly correlated with their linguistic eloquence, including clever and humorous uses of language. Human mate choice even today is often influenced by displays of cognitive abilities, through the creative use of language (e.g. Miller 2000; Franks and Rigby 2005, 208). Yes, to a large extent, human race can be seen as smart and friendly, and yet, this cannot be the whole picture. We elect politicians to govern entire countries based on how witty and skillful they are at debates, and not based on how strong or friendly they are, or indeed how good they are at solving problems. We admire and replay the memorable snippets of witty exchanges at those debates. The most eloquent speakers tend to have the highest status, even in modern societies (e.g. Locke 2009). Competition with language continues to this day, whether in more subtle ways, or through outright verbal dueling attested across cultures around the world (Locke and Bogin 2006; Locke 2009). Darwin's (1874) view was that language evolved gradually through sexual selection, and for him, language is "half art, half instinct" (634), but this creative dimension has largely been neglected in the research on language evolution.

4. **The early steps**

But how is all this relevant for the earliest steps in language evolution, when human unique cognitive abilities just started to be honed? There are proposals in the literature to the effect that some of the earliest forms of grammar, such as the first verb-noun combinations, required novelty, imagination, and quick-wittedness, in order to be useful and entertaining, and to catch an audience. In addition to many other useful functions (see below), according to the proposal in Progovac and Locke (2009) and Progovac (2015, 2016), coining compounds akin to the ones illustrated below (proxies of the earliest grammars) would have been a highly adaptive way to compete for status and sex in ancient times, when words were few, and grammars rudimentary (1 is from English, mostly taken from Weekley 1916; 2 is from Serbian, mostly taken from Mihajlović, 1992).[1]

(1) kill-joy, turn-skin (traitor), hunch-back, wag-tail, tattle-tale, scatter-brain, cut-throat, mar-wood (bad carpenter), busy-body, cry-baby, break-back, catch-fly (plant), cut-finger (plant), tumble-weed, fill-belly (glutton), lick-spit,

[1] Importantly, entrenching the abilities to coin such compositions in populations would have provided an excellent scaffolding for evolving full sentences, which feature exactly nouns and verbs (see Progovac (2015, 2016) for a reconstruction of the earliest grammars based on syntactic theory).

pinch-back (miser), shuffle-wing (bird), skin-flint (miser), spit-fire, swish-tail (bird), rattle-snake, stink-bug, tangle-foot (whiskey), tumble-dung (insect), crake-bone (crack-bone), shave-tail (shove-tail), fuck-ass, fuck-head, shit-ass, shit-head

(2) cepi-dlaka ‘split-hair’ (hair-splitter); guli-koža ‘peel-skin’ (who rips you off); vrti-guz ‘spin-butt’ (restless person, fidget); muti-voda ‘muddy-water’ (trouble-maker); jebi-vetar ‘fuck-wind’ (charlatan); vuci-guz ‘drag-butt’ (slow-moving person); poj-kurić ‘sing-dick’ (womanizer); kosi-noga ‘skew-leg’ (person who limps); podvi-rep ‘fold-tail’ (who is crestfallen); deri-muda ‘rip-balls’ (place name, a steep hill); kapi-kur ‘drip-dick’ (name of a slow water spring); plači-guz ‘cry-butt’ (crybaby)

It would have constituted an unprecedented cognitive leap to become fluent in this strategy of combining, on the spot, two very basic, concrete words in order to express a completely novel, abstract concept, often exhibiting stunning feats of metaphorical creativity. Not everybody was good at it at the time, perhaps only a few were, and those would have been the ones to pass on their genes through many generations. This initial stage of proto-grammar would have unmasked, and thus opened for selection, these otherwise latent cognitive abilities of our ancestors.[2] The successful use of such two-slot combinations would have enhanced one’s relative status first by derogating existing rivals and placing prospective rivals on notice, and second by demonstrating verbal skill and quick-wittedness (Progovac and Locke 2009). These are exactly the two types of sexual selection scenarios identified as early as in Darwin (1874): aggressive rivalry (intrasexual selection) and mate choice (intersexual selection), both also proving to be relevant in the experiment by Franks and Rigby (2005), which found that males increase their creativity with language both in the presence of attractive females and in the presence of male competitors. The compounds illustrated above are also typically humorous, invoking concrete (body) images (e.g. *scatter-brain*; *pinch-back* for a miser; *spin-butt* for a restless person; *rip-balls* for a steep hill). Humor itself is subject to sexual selection (e.g. Vrticka et al. 2013).

It is notable that the proto-grammatical compounds (1-2), serving as approximations of the earliest grammars, are not only humorous and imageable,

[2] In this respect, and in the interest of highlighting continuity, it has been reported that other primates are in principle capable of rudimentary two-slot combinations, such as *hide peanut* and *hide Kanzi* (Greenfield and Savage-Rumbaugh 1990: 161, regarding bonobo Kanzi). According to Patterson and Gordon (1993), gorilla Koko was not only capable of producing novel two-slot metaphorical combinations (e.g. ‘cookie rock,’ for a stale bun), but also of playful insult.

but also tend to express (playful) insult when referring to humans, across different cultures (Progovac and Locke 2009).[3] Paradoxically, perhaps, it is exactly in this respect that the proposal of the survival-of-the-wittiest can be cross-fertilized with the survival-of-the-friendliest proposal. As proposed in Progovac and Benítez-Burraco (2019), the gradual emergence of verbal means of competition/aggression (which certainly includes wittiness) was engaged in a feedback loop with the genetic forces working toward the reduction in physical aggression, associated with self-domestication (see also Benítez-Burraco and Progovac 2021). By affording a more adaptive (less violent) way to compete for status and sex, these items of verbal competition would have reinforced the effects of self-domestication, by providing a means for gradually replacing reactive physical aggression with verbal competition. It would have also provided a means for reducing cortisol levels through the use of verbal humor. But for this to have worked, the witty in these early stages needed to be so expressive, so creative with what little language and grammar there was at that time, to be able to outdo the eons deep exchanges and displays of physical strength. Quite an astonishing cognitive feat, and not a trivial reason at all for language to take root in populations, and to gradually grow and thrive. Crucially, when these simple verb-noun combinations started to be entrenched, they would have been used and useful for many other functions as well. In addition to naming people (in imageable, often pejorative ways), they would have also been used to name plants (*tumble-weed*) and animals (*rattle-snake*), as well as to issue commands (*Kill snake! Run Kanzi!*) or to make observations (e.g. *Snake rattle. Baby cry.*)

This approach is not (meant to be) in conflict with approaches that give primacy to pragmaticsin language evolution, rather than grammar, emphasizing continuity with other species in this respect (see e.g. Arnold & Zuberbühler, 2013; Seyfarth & Cheney, 2017; Bar-On 2021). This approach in fact reveals continuity, both when it comes to specific communicative uses of early forms of language (non-propositional) and when it comesto the earliest forms of grammar, consistent with the abilities of other primates (see footnote 2; also Progovac, 2017, for a detailed discussion). But in addition to relying on continuity, this approach also provides a point of departure, a beginning of what will turn out to be uniquely human traits. This proposal provides a fertile ground for a wide variety of hypotheses to be tested, with a goal of understanding how and why (sexual) selection, specifically for agility with language, has been shaping human nature.

[3] It seems that we are still in a proto-linguistic mindset when using this kind of imageable, derogatory language, which prefers to occur in proto-linguistic frames (for reasons and findings in this respect, see Progovac et al.'s 2018 fMRI study).

Acknowledgements

I am grateful to the three anonymous reviewers of this paper, as well as editors, who asked good questions and provided excellent starting points for further conversation and discussion.

References

Arnold, K., & Zuberbühler, K. (2013). Female putty-nosed monkeys use experimentally altered contextual information to disambiguate the cause of male alarm calls. *PLOS ONE 8*(6):e65660.

Bar-On, D. (2021). How to do things with nonwords: pragmatics, biosemantics, and origins of language in animal communication. *Biology and Philosophy, 36,* 50. https://doi.org/10.1007/s10539-021-09824-z

Benítez-Burraco, A., & Progovac, L. (2021). Language evolution: examining the link between cross-modality and aggression through the lens of disorders. *Philosophical Transactions of the Royal Society B, 376*(1824), 20200188.

Berwick R, & Chomsky N. (2016). *Why only us? Language and evolution*. London, UK: MIT Press.

Cimons, M. (2020) 'Friendliest,' not fittest, is key to evolutionary survival, scientists argue in book. *Washington Post.* ISSN 0190-8286. Retrieved 2020-10-07.

Darwin, C. (1859/1964). *On the Origin of Species*. 1964 facsimile edition. Cambridge, MA: Harvard University Press. First published in 1859.

Darwin, C. (1872). *The Expression of the Emotions in Man and Animals*. London: John Murray.

Darwin, C, M. A. (1874). *The Descent of Man, and Selection in Relation to Sex*. New edition, revised and augmented. New York: Hurst and Company.

Franks, B., & Rigby, K. (2005). Deception and mate selection: Some implications for relevance and the evolution of language. In M. Tallerman (Ed.), *Language Origins: Perspectives on Evolution* (pp. 208-229). Studies in the Evolution of Language. Oxford: Oxford University Press.

Gleeson, B.T. (2018). Masculinity and the mechanisms of human self-domestication. *BioRxiv* preprint. doi.org/10.1101/143875.

Greenfield, P. M., & Savage-Rumbaugh, E. S. (1990). Grammatical combination in Pan paniscus: Processes of learning and invention in the evolution and development of language. In S. T. Parker & K. R. Gibson (Eds.), *"Language" and intelligence in monkeys and apes: Comparative developmental perspectives* (pp. 540–578). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511665486.022

Hare, B. (2017). Survival of the Friendliest: Homo sapiens evolved via selection for prosociality. *Annual Review of Psychology, 68*(1).

Hare, B., & Woods, V. (2020). *Survival of the Friendliest: Understanding our Origins and Rediscovering our Common Humanity*. Random House, New York.

Hare, B., Wobber. V., & Wrangham, R. (2012). The self-domestication hypothesis: Evolution of bonobo psychology is due to selection against aggression. *Animal Behavior, 83.3*, 573-585.

Locke, J. L. (2009). Evolutionary developmental linguistics: Naturalization of the faculty of language. *Language Sciences, 31*, 33–59.

Locke, J. L. & Bogin, B. (2006). Language and life history: A new perspective on the evolution and development of linguistic communication. *Behavioral and Brain Science, 29*, 259–325.

Mihajlović, V. 1992. *Ime po zapovesti: Imperativni Onomastikon srpskohrvatskog jezika*. [*Name by Command*]. Belgrade: Nolit.

Miller, G. A. 2000. *The Mating Mind: How Sexual Choice Shaped the Evolution of Human Nature*. London: William Heinemann.

Patterson, F., & Gordon, W. (1993). The case for the personhood of gorillas. In P. Cavalieri and P. Singer (Eds.), *The Great Ape Project* (pp. 58–77). New York, NY: St. Martins Griffin.

Progovac, L. (2015). *Evolutionary Syntax*. Oxford: Oxford University Press. Oxford Scholarship Online.

Progovac, L. (2016). A Gradualist scenario for language evolution: Precise linguistic reconstruction of early human (and Neandertal) grammars. *Frontiers in Psychology*, *7*:1714. doi: 10.3389/fpsyg.2016.01714.

Progovac, L. (2017). Where is continuity likely to be found? Commentary on 'The social origins of language' by Robert M. Seyfarth and Dorothy L. Cheney. Edited and introduced by Michael Platt, 46-61. Princeton and Oxford: Princeton University Press.

Progovac, L., & Benítez-Burraco, A. (2019). From physical aggression to verbal behavior: Language evolution and self-domestication feedback loop. *Frontiers in Psychology*, *10*:2807. https ://doi.org/10.3389/fpsyg.2019.02807

Progovac, L., & Locke, J. L. (2009). The urge to merge: Ritual insult and the evolution of syntax. *Biolinguistics, 3.2*-3, 337-354.

Progovac, L., Rakhlin, N., Angell, W., Liddane, R., Tang, L., & Ofen, N. (2018). Neural correlates of syntax and protosyntax: An fMRI study. *Frontiers in Psychology, 9*, 2415. doi:10.3389/fpsyg.2018.02415

Seyfarth, R., & Cheney, D. (2017). The origin of meaning in animal signals. *Animal Behaviour, 124*:339–346.

Vrticka, P., Black, J.M., and Reiss, A.L. (2013). The neural basis of humor processing. *Nature Reviews Neuroscience, 14*(12):860-8. doi: 10.1038/nrn3566.

Weekley, E. (1916). *Surnames*. New York: E.P. Dutton & Co.

Wrangham, R.W. (2021). Targeted conspiratorial killing, human self-domestication and the evolution of groupishness. *Evolutionary Human Sciences*, 3, e26, page 1 of 21doi:10.1017/ehs.2021.20

Uy, J. A. C., Patricelli, G. L. & Borgia, G. (2001). Complex Mate Searching in the Satin Bowerbird Ptilonorhynchus violaceus. *The American Naturalist*, 158 (5), 530–542.
doi:*10.1086/323118*. *ISSN 0003-0147*. *PMID 18707307*. *S2CID 2506815*.

The role of audience design in an emerging language system: Evidence from a novel signaling paradigm

Casey M. Riedmann*[1] and William S. Horton[1]

*Corresponding Author: CaseyRiedmann2023@u.northwestern.edu
[1]Department of Psychology, Northwestern University, Evanston, IL, USA

The evolution of human language is informed by rapid development across many cognitive abilities. Some of these abilities have been evaluated via novel signaling experiments in which conventional signals are restricted (e.g., Sulik & Lupyan, 2018). Across four experiments, we tested whether knowing interpersonal information, operationalized here as partner age, affects signaling outcomes. We developed a word list with contemporary meanings to determine whether signalers consider age information when generating clue words for specific partners. When informed of a partner's age (i.e., 18-25 or 40-60) at the task's onset, we found some audience design evidence (Expt. 1), supporting its potential role in language evolution. When provided with frequent and disruptive reminders of a partner's age, signalers capitalized on this information more often (Expt. 2-3). Testing signalers' clues, however, did not yield more success for receivers guessing target words which were produced to be age-specific (Expt. 4).

1. Language Evolution

Humans' capacity for language is thought to have emerged within a suite of cognitive abilities (Pinker, 2010). For instance, language depends on a memory network that stores semantic information which, in turn, supports the conceptual categorization that language offers (Gong & Shuai, 2015). Such interdependencies, however, enhance the difficulty of unraveling the phylogenetic order in which various abilities evolved. Decoding this order is crucial for determining conditions that support emergence of a full-fledged language system. A recent set of ideas has focused on the importance of forming communicative intentions (Scott-Phillips, 2014). Within this context, we explore audience design as a potential prerequisite for language evolution.

1.2. Measuring Language Evolution: Novel Signaling Paradigm

Methodological innovations in experimental semiotics have contributed to new insights concerning the cognitive underpinnings of human communication (see Galantucci et al., 2012, for a review). One approach, pioneered by Scott-Phillips et al. (2009), is the *novel signaling* paradigm in which participants are tasked with conveying meaning to a partner without the ability to communicate via conventional means. This work allows insight into the conditions that could support language's emergence in an environment where conventional signals were not available. For example, with the goal of persuading a partner to move a stimulus to a target region, a conventional signal would be to simply say "move the [stimulus] to [target region]." However, when such explicit signals are forbidden (e.g., in remote contexts), conveying an intention proves significantly more difficult. However, Scott-Phillips et al. found that pairs could achieve success by assuming 'default' conventions that could function as a type of common ground.

Expanding on this work, Sulik and Lupyan (2018) explored the nature of perspective-taking in a novel signaling task. In their study, partners were assigned to either a signaler or receiver role. Signalers read a target word (e.g., *bank*) with instructions to elicit the target word from the receiver. Crucially, signalers were permitted to produce exactly one (non-target) word to signal the target. Thus, signalers could rely on conventional signals (i.e., words) in the task, but those signals were novel in use. Targets were selected for holding either symmetric or asymmetric semantic relationships with their most salient associates. The word *bank*, for example, is asymmetric in that its most frequent associate is *money*; however, *money*'s most salient associate is *cash,* <u>not</u> *bank*. To elicit a guess of *bank*, then, a better approach is to select a clue with *bank* as a strong backward associate (e.g., *teller*). This imbalance allowed the authors to investigate how signalers consider their partners' perspectives during novel signaling, reasoning that signalers who are more attuned to their partners' perspectives should produce clues that increase the likelihood of guessing the target.

In Sulik and Lupyan's (2018) initial experiment, signalers struggled to produce helpful clues for asymmetric targets. In subsequent experiments, however, the authors constrained the signal (by providing signalers with a list of clue options) or the common ground (by sharing that list with receivers), resulting in better performance. These findings suggest that, in environments where access to a conventional signal is restricted, representing an interlocutor's mental state may be an important prerequisite for communication (see Tomasello, 2008). Using this paradigm to tease out the processes that undergird humans' capacity for language invites further exploration of the various abilities that support language.

2. Testing Audience Design

Here, we adapted Sulik and Lupyan's (2018) work to test whether certain aspects of *audience design* could enhance communicative success on a linguistically restrictive task. Audience design occurs when speakers adjust utterances to a specific audience to facilitate comprehension (H. H. Clark & Murphy, 1982), and speakers frequently rely on interpersonal cues when determining their referential choices, such as whether an addressee is a native speaker (Bortfield & Brennan, 1997), or a novice or expert (Isaacs & H. H. Clark, 1987). Age is another salient cue that speakers regularly consider, tailoring speech for both children (E. V. Clark & Estigarribia, 2011) and elderly adults (Hummert et al., 1998; Kemper, 1994). Thus, knowing an interlocutor's age could influence communicators to produce age-sensitive signals.

Recognizing relevant social characteristics not only influences modern language use, but may support humans' capacity for communication more generally (Seyfarth & Cheney, 2014). Early humans evolved linguistic capabilities within social contexts in which distinguishing members and tracking individual histories was likely highly valuable. Non-human primates, for comparison, display behaviors that suggest an awareness of partner-specific history associated with enhancing fitness and offspring survival (Rosenbaum et al., 2016; Silk, 2007). Thus, individualized behaviors toward conspecifics plausibly included acts of direct communication in early humans, which could be considered precursors of audience design. Such personalized interactions reflect a broader network of cognitive abilities that allow for reasoning about others' mental states, essential for establishing a language system (Scott-Phillips, 2014).

Here, we test whether knowledge of a partner's age improves communication in a novel signaling task. Because the paradigm restricts conventional signals, signalers must forge new ways to represent their intentions, illustrating circumstances similar to ones with which early humans contended. Considering interpersonal information under these conditions would support an account of language evolution in which the ability to engage with audience design is a critical and potentially necessary feature.

2.2. *Experiment 1: Audience Design in a Novel Signaling Task*

We recruited 23 Northwestern students from an intro psychology class (avg. age: 19) to serve as signalers. Participants were monolingual English speakers who completed the task online via Qualtrics software. For each of a series of target words, signalers produced a single (non-target) clue word to elicit a guess from their partners of the target, replicating Sulik and Lupyan (2018). Participants were (falsely) informed that the partner (who did not exist) would complete the task later, so no feedback would be provided. Critically, at the task's onset, signalers were informed that the partner was a native English speaker from the U.S. belonging to one of two age groups: 18-25 or 40-60. The partner's age group was randomly assigned between-subject with the expectation that signalers would produce clues tailored for someone in that age range.

Signalers were presented with a list of 40 words in random order. Upon reading each word, they were asked to type a clue word into a textbox for their partner. Twenty of the 40 words were used by Sulik and Lupyan (2018)—i.e., words with either symmetric or asymmetric primary associates—and the other 20 were critical words that held "contemporary" meanings. These contemporary words are recently popularized homonyms (e.g., *tea*) which carry both classic meanings (e.g., the beverage) and modern meanings that have recently emerged in youth vernacular (e.g., gossip). Our goal was to test whether signaler behavior differed between the two partner conditions, predicting that those assigned to older partners would be more likely to produce clues with traditional meanings, while those assigned to younger partners would be more likely to produce clues with contemporary meanings, a pattern indicative of audience design. Moreover, we expected this pattern because our participants were peers to younger partners, suggesting a loosely shared inventory of cultural knowledge, including neologisms.

Sulik and Lupyan (2018) analyzed responses by evaluating clues' forward and/or backward association strength relative to the target based on Nelson et al.'s (2004) Free Association Norms corpus. However, contemporary meanings were unavailable in existing norms and responses could not be analyzed by association strength. Rather, we coded clues categorically as either traditional, contemporary, or some other meaning[1]. If the partner's purported age range had any bearing on these responses, then the contemporary words should have been more likely to elicit a contemporary response for younger partners than for older partners.

Overall, clue word production suggested that signalers engaged in audience design. A mixed effect logistic regression assessed the likelihood of producing a contemporary clue given a contemporary target[2]. The model revealed an effect of partner's age on clue production, with signalers assigned to younger partners producing contemporary clues for critical targets more often (17.9% of trials) than those assigned to older partners (12.7%), $b = 0.37$, $p = .049$. This effect suggests that access to a partner's age influences the production of signals in a novel signaling task.

2.3. *Experiment 2: Boosting the Effect of Audience Design*

In Expt. 2 we enhanced the salience of the partner's age to account for participants who may have overlooked age in Expt. 1 (which appeared only at the task's onset). We recruited 30 more undergraduates (avg. age: 18.33) and replicated Expt. 1's materials and procedure. The critical change in Expt. 2 was providing signalers with continuous access to partners' demographic information, On-screen reminders appeared alongside each target: "Reminder: Your partner is [18-25/40-60] years old, speaks English, and resides in the U.S." The age reminder corresponded to participant's randomly assigned age condition.

On-screen age information enhanced signalers' engagement of audience design relative to Expt. 1. Analyzing the critical contemporary words, a mixed effect logistic regression model revealed that signalers produced more clues corresponding to contemporary meanings when assigned to a younger partner (22%) than an older partner (13%), $b = 0.49$, $p = .001$; suggesting that continuous access to age encourages audience design more explicitly in contexts where conventional signals are restricted.

---

[1] Sulik and Lupyan's original words were not necessarily polysemous, so the analysis procedure only applies to contemporary words (which are of theoretical interest). Evaluating association strength for the original words, however, supports S&L's findings, a replication providing reliability for the paradigm.
[2] For Expt. 1-3, no trials/participants were removed prior to analysis; however, given the limited sample size, the full mixed effect model could not converge. The results reported here are from reduced versions of the model in which higher order effects were systematically removed until the model fit. In the summary (below; see Table 1), random effects are reported for the data aggregated across Expt. 1-3.

2.4. *Experiment 3: Audience Design Across Multiple Interlocutors*

In Expt. 3, we replicated our procedure using a within-subject design wherein signalers (n=26; avg. age: 19.08) provided clues to receivers from *both* age groups. The task occurred in-lab, though all other procedures were preserved. Participants were (randomly) assigned to a partner in one age group, and then halfway through the task informed that the partner had switched to someone in the other group, making age more evident. Contemporary targets were balanced across both partners.

Like Expt. 1 and 2, Expt. 3 continued signalers' trend of increasing influence of partner age as age was made more salient. A logistic mixed effect model supported the idea that addressing a younger partner produced more contemporary clues (27.7%) than an older partner (14.2%) for contemporary targets, $b = 0.75$, $p < .001$. Overall, Expt. 3 finds that switching partners fosters greater signal differentiation in the task.

2.5 *Summary: Experiments 1-3*

Taken together, Expt. 1-3 demonstrate audience design in novel signaling tasks with clues tailored to signaler's partner's age range. Evidence of audience design grew as age was made more explicit between experiments. To address sample size concerns, we implemented a logistic mixed effect model using data pooled across all three experiments (in which the basic procedure was the same). Participants produced 3-4 contemporary clues on average, with only 1 participant failing to produce any. The model revealed an effect of partner age, with contemporary meanings more likely to be produced for younger partners (22.6%) than older ones (13.3%), $b = 0.56$, $p < .001$. All results are summarized in Table 1:

Table 1. A summary of the logistic mixed effect models from Expt. 1-3

	N	*Odds Ratio*	*CI*	τ_{00} *(subj.)*	τ_{00} *(item)*	*ICC*
Expt. 1	23	1.44	1.00-2.07	0.05	10.90	0.77
Expt. 2	30	1.64	1.22-2.20	0.08	4.59	0.59
Expt. 3	26	2.12	1.54-2.93	0.03	8.17	0.71
Expt. 1-3	**79**	**1.76**	**1.45-2.13**	**0.13**	**6.52**	**0.67**

2.6. *Experiment 4: Receiving Signals*

In Expt. 4, we tested whether the signals (most frequent responses from Expt. 1-2, whether contemporary or not) were actually effective at eliciting guesses of the target. We recruited 47 students (avg. age: 18.77) to be receivers, randomly assigned (between-subject) to read clues initially generated for either a younger or older partner. Because receivers matched the younger receivers' (and original signalers') age range, it was expected that clues generated for younger partners may prove more successful. The task occurred in-lab with receivers viewing clues on a computer and being prompted to guess the original target word. Participants were informed that clues had been generated by a previous signaler.

The task proved difficult overall (avg. success rate = 27.8%). A logistic mixed effect model revealed no evidence that receivers performed better given clues that were initially generated for a younger partner (27.2%) than an older one (28.3%), b = -0.04, p = 0.75. Thus, while Expt. 1-3 revealed evidence of audience design in a novel signaling task, the top clues produced in Expt. 1-2 were only moderately successful, with no apparent effect between clues generated for younger and older partners.

3. Conclusion

In Expt. 1-3, we find that signalers engaged in audience design by incorporating age-specific information during non-conventional signaling. Age information played a larger role in signaling when it was made more explicit, as the trends between experiments shows. Despite support for audience design, though, Expt. 4 reveals that clues tailored to younger partners did not improve success when shared with younger receivers. Still, access to interpersonal information appears to influence signaling when conventional signals are restricted. In a prelinguistic society, awareness of social information about one's partner could serve as a catalyst for conveying meaning. For instance, conveying a communicative intention may be realized by an enhanced understanding of an interlocutor's mental state and sociocultural background. However, there are limits to any far-reaching conclusions on this point. Specifically, this paradigm demonstrates how signalers generate alternatives when they lack access to conventional signal. In some aspects, this is a valid simulation of developing abstract linguistic representations *de novo*, but it is also limited in the sense that signalers (and their partners) have preexisting access to a fully developed lexicon. Nevertheless, when prevented from utilizing lexical knowledge conventionally, signalers could compensate by taking advantage of other cognitive skills, including access to stereotypical knowledge about the likely idiolects of individuals of a certain age range. Although audience design enhances standard language production, speculation about its importance in language's emergence remains plausible though ultimately unresolved.

References

Bortfield, H., & Brennan, S. E. (1997). Use and acquisition of idiomatic expressions in referring by native and non-native speakers. *Discourse Processes*, *23*(2), 119-147.

Clark, E. V., & Estigarribia, B. (2011). Using speech and gesture to introduce new objects to young children. *Gesture*, *11*(1), 1-23.

Clark, H. H., & Murphy, G. L. (1982). Audience design in meaning and reference. In *Advances in Psychology* (Vol. 9, pp. 287-299). North-Holland.
Galantucci, B., Garrod, S., & Roberts, G. (2012). Experimental semiotics. *Language and Linguistics Compass*, *6*(8), 477-493.
Gong, T., & Shuai, L. (2015). Modeling coevolution between language and memory capacity during language origin. *Plos one*, *10*(11), e0142281.
Hummert, M. L., Shaner, J. L., Garstka, T. A., & Henry, C. (1998). Communication with older adults: The influence of age stereotypes, context, and communicator age. *Human Communication Research*, *25*(1), 124-151.
Isaacs, E. A., & Clark, H. H. (1987). References in conversation between experts and novices. *Journal of experimental psychology: general*, *116*(1), 26.
Kemper, S. (1994). Elderspeak: Speech accommodations to older adults. *Aging and Cognition*, *1*(1), 17-28.
Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402-407.
Pinker, S. (2010). The cognitive niche: Coevolution of intelligence, sociality, and language. *Proceedings of the National Academy of Sciences*, *107*, 8993-8999.
Rosenbaum, S., Hirwa, J. P., Silk, J. B., & Stoinski, T. S. (2016). Relationships between adult male and maturing mountain gorillas (*Gorilla beringei beringei*) persist across developmental stages and social upheaval. *Ethology*, *122*(2), 134-150.
Scott-Phillips, T. (2014). *Speaking our minds: Why human communication is different, and how language evolved to make it special*. Bloomsbury Publishing.
Scott-Phillips, T. C., Kirby, S., & Ritchie, G. R. (2009). Signalling signalhood and the emergence of communication. *Cognition*, *113*(2), 226-233.
Seyfarth, R. M., & Cheney, D. L. (2014). The evolution of language from social cognition. *Current Opinion in Neurobiology*, *28*, 5-9.
Silk, J. B. (2007). Social components of fitness in primate groups. *Science*, *317*(5843), 1347-1351.
Sulik, J., & Lupyan, G. (2018). Perspective taking in a novel signaling task: Effects of world knowledge and contextual constraint. *Journal of Experimental Psychology: General*, *147*(11), 1619.
Tomasello, M. (2010). *Origins of human communication*. MIT press.

The evolution of phonological dispersion: New experimental results

Gareth Roberts[*1] and Robin Clark[1]

[*]Corresponding Author: gareth.roberts@ling.upenn.edu
[1]Linguistics Department, University of Pennsylvania, Philadelphia, USA

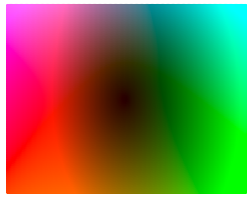
Phonological inventories seem to exhibit greater structure than expected by chance, with dispersion in vowel spaces being a well-known example (De Boer, 2000; Lindblom, 1986; Lindblom & Maddieson, 1988). How does such structure emerge and evolve? One hypothesized explanation is that dispersed systems are a response to pressures acting on perceivers—dispersal aiding distinctiveness—and on producers—with articulations at the edges of the space being easier to produce reliably (Liljencrants & Lindblom, 1972; Schwartz, Boë, Vallée, & Abry, 1997). However, the space where such dynamics play out is not uniform, and we might expect a number of factors to modulate the process, including noise, and transitions between units in production (De Boer, 2016; Carré, 2009). This account is not specific to language, suggesting that the same factors should lead to the emergence of similar structure in non-linguistic communication systems.

Roberts and Clark (2020, 2023) investigated this by having pairs of participants play a computer game in which they took turns to communicate silhouettes of animals using colors. The sender on a given turn moved a finger around on a trackpad to select series of colors from a continuous underlying colorspace. The structure of the colorspace was manipulated to vary whether the most reliably locatable areas of the trackpad for the sender lined up with the most distinct colors. They found higher dispersion than would be expected by chance, driven in large part by perceptual demands. Communicative success was lower when production and perception demands were less aligned. Roberts and Clark (2023) analyzed the process by which dispersion came about, finding it was not planned from the start but emerged as a consequence of small-scale choices and adjustments over time.

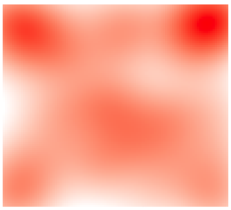
We replicated Roberts and Clark (2020, 2023) with several changes ($N = 160$). First, the colorspace was redesigned so that,in one condition, distinct colors were available throughout the space, reducing bias either for or against dispersion (Fig. 1a–1b).[1] Second, we manipulated the presence of noise, operationalized as random color deviation. Third, we manipulated the minimum number of colors (1 vs. 2) that a signal had to contain. All conditions were crossed.

[1]NB: Colors in image may seem more indistinguishably dark than reality if viewed from a distance.

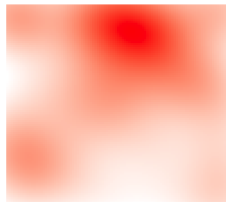
(a) Light-center colorspace



(b) Dark-center colorspace[1]



(c) Heatmap for light-center, high-noise, 2+ color condition



(d) Heatmap for dark-center, low-noise, 1+ color condition

Figure 1. Example colorspaces and heatmaps of signal-initial selections.

We measured dispersion as mean pairwise distance between signal units. This was significantly greater than chance across conditions, $\beta = 0.074, SE = 0.02, t = 4.52, p < 0.001$ It was greatest in the light-center, noise, and 2+ minimum-length conditions. Fig. 1c–1d shows heatmaps of coordinates for color selections for maximally and minimally dispersed combinations of conditions. A linear model with mean pairwise distance between signal coordinates as DV, noise and minimum signal length as predictors, along with interaction terms, did not find a significant effect ($p > 0.6$). Signal-initial colors were in fact more extreme than later colors, but this within-signal effect seems to have evened out over the inventory as a whole. The lack of an effect of noise is likely due to the communication medium already being sufficiently noisy that participants were responding to noise across all conditions.

A mixed model with mean pairwise distance as dependent variable and colorspace as predictor, with random intercepts for noise and minimum signal length, found a significant effect: $\beta(74) = 0.06, SE = 0.02, t = 2.79, p < 0.01$. These results (taken together with earlier work) suggest that phonological dispersion may be best explained as resulting from communicative pressures interacting with the topology of the signaling space (cf. Schwartz et al., 1997).

Acknowledgements

We thank many members of the Cultural Evolution of Language Lab for running trials, and we gratefully acknowledge funding from the National Science Foundation (award number 1946882).

References

Carré, R. (2009). Dynamic properties of an acoustic tube: Prediction of vowel systems. *Speech Communication*, *51*(1), 26–41.

De Boer, B. (2000). Self-organization in vowel systems. *Journal of Phonetics*, *28*(4), 441–465.

De Boer, B. (2016). Modeling co-evolution of speech and biology. *Topics in Cognitive Science*, *8*(2), 459–468.

Liljencrants, J., & Lindblom, B. (1972). Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language*, *48*(4), 839–62.

Lindblom, B. (1986). Phonetic universals in vowel systems. *Experimental phonology*, 13–44.

Lindblom, B., & Maddieson, I. (1988). Phonetic universals in consonant systems. *Language, Speech, and Mind*, 62–78.

Roberts, G., & Clark, R. (2020). Dispersion, communication, and alignment: An experimental study of the emergence of structure in combinatorial phonology. *Journal of Language Evolution*, *5*(2), 121–139.

Roberts, G., & Clark, R. (2023). The emergence of phonological dispersion through interaction: An exploratory secondary analysis of a communicative game. *Frontiers in Psychology*, *14*, 1130837.

Schwartz, J.-L., Boë, L.-J., Vallée, N., & Abry, C. (1997). The dispersion-focalization theory of vowel systems. *Journal of phonetics*, *25*(3), 255–286.

Gender balance in evolutionary linguistics

Seán G. Roberts*[1], Christine Cuskley[2]

*Corresponding Author: RobertsS55@cardiff.ac.uk
[1] School of English, Communication and Philosophy, Cardiff University, Cardiff, UK
[2] Language Evolution, Acquisition and Development Group, Newcastle University, Newcastle Upon Tyne, UK

1. Introduction

One important step towards improving equity, diversity and inclusion in a given field is to be aware of current imbalances in the distributions of identities of researchers. However, intuitions about these imbalances can be inaccurate due to various cognitive biases about the perception of individuals and the norms in the field (e.g. García-González, Forcén & Jimenez-Sanchez, 2019). This complicates decision making about where to invest resources in student recruitment, job recruitment and outreach. One productive step is to explicitly monitor imbalances using objective methods. This study looks at the distribution of genders across sub-fields of evolutionary linguistics, broadly construed.

2. Methods

For the review stage, subfields were defined according to categories and keywords used in the Evolution of Language conferences (e.g. acquisition, phylogenetics, sign language), as identified in Wacewicz et al., (2022). Web of science was used to find hundreds of journal papers about language evolution from each sub-field published within the last 10 years. This was done by searching for field-level descriptors ("language evolution", "evolution of language", "cultural evolution") together with a specific sub-field descriptor (e.g. "sign language"). The 100 authors with the highest number of publications in each sample were identified. Each author was manually coded for conferred gender based on academic profiles, using the methods from Cuskley et al. (2020) and Rennick et al. (2023). The distribution of genders in each language evolution sub-field was calculated and compared to the distribution of genders in the broader sub-field outside of language evolution. The same method was then repeated for each subfield in general by searching only for the subfield descriptors and omitting language evolution keywords.

3. Results

Figure 1 shows a sample of the results. The subfield of phylogenetics had the lowest proportion of female authors writing about language evolution (15%) and the field of sign language had the highest (49%). These proportions were roughly equal to the proportion of female authors in the general literature. However, papers on language evolution and language acquisition had half the proportion of female authors compared to their general field.

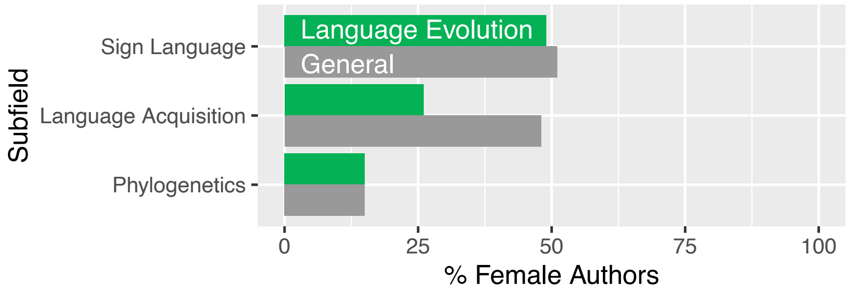


Figure 1. Sample of results: Proportion of female authors in papers published in evolutionary linguistics (green) and in general (gray) in three subfields.

4. Discussion

There appear to be several types of bias in the results. Some fields like linguistic phylogenetics may have low proportions of female authors because the general field has low proportions, possibly stemming from general longstanding biases in their feeder subjects like computer science and biology (Huang et al., 2020). In contrast, the gap for language acquisition may be due to factors specific to language evolution. For example, the historical baggage that comes with the necessary theoretical commitments to evolutionary theory may be perceived as an ethical barrier, and women may be less willing to engage with this than men (Kennedy & Kray, 2014). Alternatively, men may be more willing to do research outside their core field than women, though some studies show the opposite pattern (Pinheiro, 2022). Finally, there may be gaps between the perception of balance in a subfield and the actual distribution. For example, sign language may be perceived to be dominated by women (e.g. the majority of researchers who presented on sign language at EvoLang are female, see also e.g. MacDougall et al., 2012), but the publication data is more balanced. In addition, seniority and sampling biases need to be accounted for. Understanding these patterns is key to ensuring equitable access to the field of evolutionary linguistics.

Acknowledgements

SGR was supported by an AHRC grant AH/T006927/1.

References

Cuskley, C., Roberts, S. G., Politzer-Ahles, S., & Verhoef, T. (2020). Double-blind reviewing and gender biases at EvoLang conferences: An update. *Journal of Language Evolution*, *5*(1), 92-99.

MacDougall, Diana E. M.A., CI and CT (2012) "Gendered Discourse and ASL-to-English Interpreting: A Poststructuralist Approach to Gendered Discourse and the ASL-to-English Interpretive Process," *Journal of Interpretation*: Vol. 19: Iss. 1, Article 2.

García-González, J., Forcén, P., & Jimenez-Sanchez, M. (2019). Men and women differ in their perception of gender bias in research institutions. *PloS one*, *14*(12), e0225763.

Huang, J., Gates, A. J., Sinatra, R., & Barabási, A. L. (2020). Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proceedings of the National Academy of Sciences*, 117(9), 4609-4616.

Kennedy, J. A., & Kray, L. J. (2014). Who is willing to sacrifice ethical values for money and social status? Gender differences in reactions to ethical compromises. *Social Psychological and Personality Science*, 5(1), 52-59.

Rennick, S., Clinton, M., Ioannidou, E., Oh, L., Clooney, C., Healy, E., & Roberts, S. G. (2023). Gender bias in video game dialogue. *Royal Society Open Science*, *10*(5), 221095.

Wacewicz, S., Sibierska, M., Placiński, M., Szczepańska, A., Poniewierska, A., Ng, Y., & Żywiczyński, P. (2022). The scientometric landscape of Evolang: A comprehensive database of the Evolang conference. *Journal of Language Evolution*, *7*(2), 246-25.

Cross-population variation in usage of a call combination: evidence of signal usage optionality in wild bonobos

Isaac Schamberg [*1,2,3,4], Martin Surbeck[3,5], and Simon Townsend[1,2,6]

[*]Corresponding Author: isaac.schamberg@gmail.com
[1] Evolutionary Anthropology, University of Zurich, Switzerland
[2]Center for the Interdisciplinary Study of Language Evolution (ISLE), University of Zürich, Switzerland
[3]Human Evolutionary Biology, Harvard University, USA
[4]School of Anthropology and Museum Ethnography, University of Oxford, UK
[5] Evolutionary Anthropology, Max Planck Institute, Germany
[6]Psychology, University of Warwick, UK

The relationship between a signal's form and its function is foundational to all systems of communication and profoundly influences a communication system's expressive potential. In language, the relationship between a word's sound and its meaning is said to be 'arbitrary' because the association is a matter of socio-linguistic convention, rather than an obligatory or natural connection. Such arbitrariness is one of the key design features responsible for language's extreme lability and adaptability. Understanding arbitrariness, and its evolution, therefore, is essential in any account of the evolution of language. To shed light on the phylogeny of the phenomenon, it is necessary to take a comparative approach and examine arbitrariness (and related capacities) in the communication of non-human animals.

Non-human communication systems do not appear to exhibit the degree of arbitrariness present in language, but the precise connection between signal and function (or meaning) in animal communication is an open question. Several studies have challenged the notion that arbitrariness is unique to language by documenting changes in call structure across time (Mitani and Gros-Lous 1998; Crockford et al. 1994; Watson et al. 2015), and developmental functional flexibility (Dezecache et al. 2021). The extent to which these examples of 'signal adjustment optionality' (*sensu* Watson et al. 2022), is mirrored by a similar capacity for 'signal usage optionality' (*sensu* Watson et al. 2022) is largely unknown (but see Lameira et al. 2013 for evidence of 'signal usage optionality' in Borean orangutans).

We address this question by comparing the usage of long-distance vocalizations produced in two populations of bonobos (*Pan paniscus*). Previous work has demonstrated that two long-distance signals–high hoots (HHs) and the whistle-high hoot combination (W+HHs) are associated with distinct patterns of behavior and likely have different functions from one another (Schamberg et al. 2016; 2017). Here, we present data on the contexts in which HHs and W+HHs are produced in order to investigate potential shifts in call usage between populations.

Data for this study were collected at two field sites: LuiKotale and Kokolopori. Subjects (n=19 at Luikotale, n=32 at Kokolopori) were followed on foot and vocalizations were recorded with a directional microphone. Observers recorded HH and W+HH, and subsequently assigned each utterance to one of the following contexts: travel, arrival, feeding, or rest.

At both sites, bonobos produced W+HHs in all four contexts, but the predominant context accompanying call production differed between the two populations. At Kokolopori, the majority (22/42) of W+HHs were produced upon arrival at a fruiting tree. At LuiKotale, a plurality (20/52) of W+HHs were produced while resting. Overall, W+HH production contexts differed significantly between the two populations (full-null comparison: df= 3, $\chi^2 = 20.67$, $p< 0.001$).

Subjects produced a majority of HHs during periods of feeding or resting (75/95 at LuiKotale and 37/51 at Kokolopori), and there was no significant difference between HH contexts in the two populations (full-null comparison: df=3, $\chi^2 = 4.311$, $p=0.230$)

Our results reveal a between-population difference in bonobos' use of the W+HH call combination. Bonobos at the Kokolopori field site were significantly more likely to produce W+HHs upon arrival at a feeding tree, compared to bonobos at the LuiKotale field site. In contrast, we found no difference in the usage of HHs between the two populations. The contrasting findings regarding usage of HHs and W+HHs indicate that the shift in W+HH usage observed between LuiKotale and Kokolopori does not reflect a broader change in activity budgets related to socio-ecological factors; rather, the difference in W+HH usage may represent an example of signal usage optionality–i.e., bonobos in the two populations may use the same signal for subtly different purposes.

Acknowledgements

We would like to thank the LuiKotale Bonobo Project for offering access to the study site, and the people of Lompole village for hosting researchers in their forest. We would also like to thank to thank the Institut Congolais pour la Conservations de la Nature and the Ministry of Scientific Research and Technology in the Democratic Republic of the Congo for their support and permission to work in the Kokolopori Bonobo Reserve and Salonga National Park in Democratic Republic of Congo.

References

Crockford, C., Herbinger, I., Vigilant, L., & Boesch, C. (2004). Wild chimpanzees produce group-specific calls: a case for vocal learning?. *Ethology*, *110*(3), 221-243.

Dezecache, G., Zuberbühler, K., Davila-Ross, M., & Dahl, C. D. (2021). Flexibility in wild infant chimpanzee vocal behavior. Journal of Language Evolution, 6(1), 37-53.

Mitani, J., & Gros-Louis, J. (1998). Chorusing and call convergence in chimpanzees: Tests of three hypotheses. *Behaviour*, *135*(8), 1041-1064.

Lameira, A. R., Hardus, M. E., Nouwen, K. J., Topelberg, E., Delgado, R. A., Spruijt, B. M., ... & Wich, S. A. (2013). Population-specific use of the same tool-assisted alarm call between two wild orangutan populations (Pongopygmaeus wurmbii) indicates functional arbitrariness. PLoS One, 8(7), e69749.

Schamberg, I., Cheney, D. L., Clay, Z., Hohmann, G., & Seyfarth, R. M. (2016). Call combinations, vocal exchanges and interparty movement in wild bonobos. *Animal Behaviour*, *122*, 109-116.

Schamberg, I., Cheney, D. L., Clay, Z., Hohmann, G., & Seyfarth, R. M. (2017). Bonobos use call combinations to facilitate inter-party travel recruitment. *Behavioral ecology and sociobiology*, *71*, 1-8.

Watson, S. K., Townsend, S. W., Schel, A. M., Wilke, C., Wallace, E. K., Cheng, L., ... & Slocombe, K. E. (2015). Vocal learning in the functionally referential food grunts of chimpanzees. *Current Biology*, *25*(4), 495-499.

Watson, S. K., Filippi, P., Gasparri, L., Falk, N., Tamer, N., Widmer, P., ... & Glock, H. J. (2022). Optionality in animal communication: a novel framework for examining the evolution of arbitrariness. *Biological Reviews*, *97*(6), 2057-2075.

Great ape interaction: Ladyginian but not Gricean

Thom Scott-Phillips* and Christophe Heintz [2]

* Corresponding Author: thom.scottphillips@gmail.com
[2] Department of Cognitive Science, Central European University, Vienna

Non-human great apes inform one another in ways that can seem very humanlike. At the same time, there are also some manifest differences. How to account for these similarities and differences in a unified way remains a major challenge. Here we précis our recent analysis (Scott-Phillips & Heintz, 2023a). We make a key distinction between the expression of intentions (Ladyginian) and the expression of specifically informative intentions (Gricean). We hence distinguish several varieties of meaning that are continuous with one another. We conclude that the origins of linguistic meaning lie in gradual changes in social cognition, and in communication systems.

Many advances in understanding great ape interaction have been achieved in the past 40 or so years (Byrne et al., 2017; Tomasello & Call, 2019). It is now clear that human modes of interaction are not wholly apart from those of other great apes (hereafter: great apes). At the same time, there remain some manifest differences, most obviously the enormous range and scope of human expression.

In recent years a consensus has emerged that further progress requires scratching beneath the surface: asking what computational tasks deliver observed behaviours (e.g. Graham et al., 2020; Heesen & Fröhlich, 2022; Warren & Call, 2022; Heintz & Scott-Phillips, 2023). Furthermore, existing computational descriptions of human interaction do not include many gradations, which limits the utility of cross-species comparisons. So for deeper understanding we need a framework for interaction that specifies computational tasks and allows for gradations.

Here we present (in brief) a new analytical framework for the cognitive description of interaction. We distinguish in particular expression of intentions from expression of specifically informative intentions. We use this distinction to differentiate some varieties of 'meaning' that are continuous with one another, in contrast to the dichotomy of 'natural' and 'non-natural' meaning. Further detail appears in newly published research (Scott-Phillips & Heintz, 2023a).

Layers of attention manipulation

Figure 1 summarises our 'special case of' framework for classifying different modes of the intentional manipulation of attention. These distinctions will allow us to describe great ape gesture in a way that both recognises its cognitive sophistication, and also accounts for observable differences with humans.

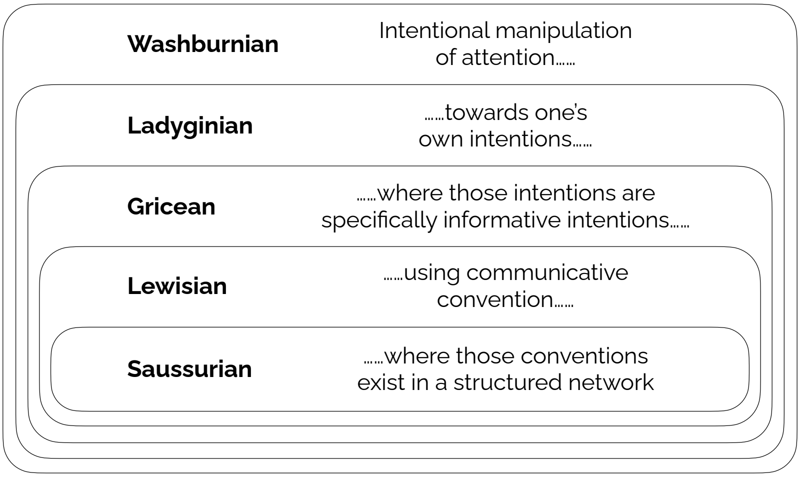


Figure 1: *Graded distinctions in modes of attention manipulation*. The distinctions between these subsets are graded rather than categorical, and shifts between them are gradual. These shifts are critical to the origins of linguistic meaning.

The outermost subset includes all instances of the intentional manipulation of attention. This is effectively how the concept of 'intentional' expression has mostly been used in comparative cognition. We label this subset *Washburnian* after Margaret Floy Washburn, who argued that the difference between the human and the non-human psyche was a difference of degree, and not a difference in kind (Washburn, 1908).

In the next subset, individuals intentionally manipulate others' attention towards evidence of their (the focal individual's) own intentions: which could be to play, to travel, to have sex, to be groomed, and so on. This is possible if the target audience has social cognitive capacities able to identify others' intentions. We label this subset *Ladyginian* after Nadezhda Ladygina-Kohts (born Nadezhda Ladygina), who was an early pioneer in the comparative study of great ape social cognition (Ladygina-Kohts & de Waal, 2002).

In the third subset, individuals intentionally manipulate others' attention towards evidence of a specific type of intention, namely informative intentions. Eating, for instance, is an intentional behavior in humans; but sometimes humans eat in an elaborated or slightly exaggerated way, perhaps accompanied by facial expressions, to suggest to others that the food is tasty, revolting, generous, or fancy. When we do this, we have a specifically informative intention that the audience learns something about the food, and we satisfy this intention by providing evidence of it i.e. by providing evidence of the intention itself. Such behaviors are commonly called *Gricean* after the philosopher Paul Grice, who developed the idea that meaning in human interaction derives from the provision of evidence for informative intentions (1957, 1989). The labels 'interactional engine' and 'ostensive communication' are often used in ways roughly synonymous with how we use 'Gricean' here (e.g. Origgi & Sperber, 2000; Tomasello, 2008; Scott-Phillips, 2015; Levinson, in press).

So the difference between Ladyginian and Gricean behavior is that whereas Ladyginian behavior intentionally reveals an intention, Gricean behavior intentionally reveals specifically informative intentions. This is not a behavioral distinction but a cognitive one. Both entail informative intentions; the difference is how the informative intention is satisfied. With Ladyginian behavior it is satisfied by making manifest the embedded intention ('I want to play'), while with Gricean behavior it is satisfied by making manifest the informative intention itself ('I want you to believe that I want to play'). This may seem more 'elaborate', but this does not mean it is cognitively 'demanding' or 'complex'. The distinction between Ladyginian and Gricean is, we believe, essential for current and future understanding of great ape interaction.

The fourth and fifth subsets include cases where Gricean behavior is performed by the use of particular, culturally evolved tools. The fourth subset, *Lewisian* (after David Lewis), includes conventions such as used in nodding, pointing or shrugging. The fifth subset, *Saussurian* (after Ferdinand de Saussure), includes cases where the conventions are (self-)organized in highly structured networks. These networks are commonly called 'languages'.

Great ape interaction is (at least) Ladyginian

We focus on the gestural domain. Gestures are certainly not the only modality of great ape interaction—vocalizations and facial expression are also important—but two related features of gesture make it a suitable focus for detailed analysis. First, it is the domain where the evidence for cognitively rich behavior is most compelling and uncontroversial (but see e.g. Crockford et al., 2017). Second, it is also where there has been greater dedicated research attention.

Over the past 15 years, a research agenda that in effect directly targets Ladyginian behavior has proven fruitful and productive (e.g. Hobaiter & Byrne, 2014; Byrne et al., 2017). The main research innovation has been to focus on 'apparently satisfactory outcomes': to observe and measure what reactions cause gesturers to cease gesturing. A large array of distinct gestures have been identified in this way. Exactly how many depends on details of definition and granularity, but there are certainly scores of them, and many seem to be common across great ape species.

We suggest that these empirical successes are evidence that great ape gesture has a Ladyginian character, because the focus on 'apparently satisfactory outcomes' effectively targets Ladyginian behavior directly. It asks, 'What intentions did the gesturer reveal, which have now been satisfied by the audience?'. The label 'Ladyginian' has not (yet) been used to describe this approach, but Ladyginian behaviors are, we suggest, what has been targeted, and what has hence led to considerable empirical successes.

The key question is whether great ape gesture is also Gricean.

Great ape interaction is (probably) not Gricean

Distinguishing Gricean and Ladyginian modes of interaction is challenging from a methodological point of view, for at least three reasons. First, the distinction is cognitive rather than behavioral: what differentiates Gricean from Ladyginian modes of interaction is not any specific behavior, but the underlying cognitive processes from which behaviors derive. Second, both Gricean and Ladyginian modes of interaction entail satisfying an informative intention. They differ just in how the informative intention is satisfied (see above). Third, both Gricean and Ladyginian modes of interaction are context sensitive: communicators must be sensitive to what audiences can perceive and infer, and take this into account in their expressive behavior.

Nevertheless, Gricean and Ladyginian modes of interaction can be distinguished empirically. (a) On the production side, one approach is to contrast the different reactions that Gricean producers and Ladyginian producers should expect from others. (b) Another production-side approach is to motivate behaviors that are only possible among Gricean individuals. (c) On the audience side, results from several experimental tasks suggest—tentatively at least—that great apes do not ordinarily seem to expect communicators to be Gricean.

We do not have space to review all these possibilities here. Rather, we highlight one example of type (c). Audiences that expect communicators to be Gricean should show a strong sensitivity to the audience's prior knowledge about the communicator's knowledge. One suitable test would be experiments in which the independent variable is the audience's knowledge of the communicator's knowledge, and the dependent variable is the audience's reaction to communicative stimuli. Human infants show differential responses in these two conditions (Tauzin & Gergely, 2018), but there is no similar demonstration in any great ape species.

The claim that great ape interaction is Ladyginian but not Gricean does not preclude the possibility that some of the cognitive capacities necessary for Gricean interaction could, in principle, emerge in great apes living in conditions of enculturation. However, there is a difference between, on the one hand, the presence of a cognitive capacity in the ordinarily developing phenotype of a species; and, on the other, the emergence of a cognitive capacity in specific individuals by virtue of individual experience. Thus, we are not suggesting is that the cognitive capacities for Gricean interaction are wholly impossible in great apes. We are observing that, if they are present, they are still unspecialised, disfluent, not a regular part of the environment, and not part of the ordinarily developing phenotype. This is all in contrast to humans, where the relevant capacities are part of the ordinarily developing phenotype. In short, only humans are 'natural Griceans'. Other great apes appear to be 'natural Ladyginians'.

This conclusion is potentially convergent with some other analyses (e.g. Gómez, 1994; Moore, 2017; Geurts, 2022; Warren & Call, 2022), but those other analyses do not make a clear distinction between Ladyginian and Gricean modes of interaction. We are arguing this distinction is crucial for understanding both similarities and differences between humans and other great apes.

Ladyginian description & analysis

Here we reinterpret one example of great ape interaction captured on video (originally from Fröhlich et al., 2016). We aim to show, briefly, how the concept of Ladygianian behavior enriches understanding of the natural phenomena. We shall discuss this example in more detail in the conference presentation.

A mother makes multiple attempts to initiate travel with her infant across a water pond. Between 00:03 and 00:08 she pulls on a branch that the infant is sitting on. Making the assumption that great ape gesture is Ladyginian, we interpret this as the intentional expression of an intention to travel. The gesture is not successful: the infant does not move. At 00:15 the mother moves closer to the infant and briefly puts her (the mother's) hand on her own back. The method of observing 'apparently satisfactory outcomes' (see above) identifies this as a commonly used gesture, named 'Present Climb On' to describe its common usage: to present a body part onto which the audience is expected to climb (Hobaiter & Byrne, 2014). We interpret this momentary gesture as the intentional expression of an intention that the infant climb upon the mother's back, so that they can travel. This also fails—the infant remains unresponsive—and so between 00:20 and 00:26 the mother reverts to the original strategy of pulling on the branch the infant is sitting on. When this too fails, the mother travels back across the pond. Now closer, the infant finally climbs onto her back.

Varieties of meaning

Paul Grice famously distinguished 'natural meaning' from 'non-natural meaning' (1957). This is a binary distinction, and as such does not lend itself to gradualism. Our 'special case of' framework (Figure 1) helps to distinguish some notions of meaning that are continuous with one another, and hence sketch some gradations that are highly relevant for cross-species comparisons.

In the most general sense, 'meaning' is a property of a relationship between an item and a cognitive system, such as an individual mind. Anything can 'have' meaning, just so long as it is processed by (or 'is informative for') some cognitive system. But only a subset of the many possible sources of meaning in the world derive from an individual's intention to have a cognitive effect on others. Specifically:

- In cases of meaning$_W$ (Washburnian), individuals intend to act on others' mental states but do not have any particular motive to reveal this intention. Many clothing choices, for instance, are made to express certain attitudes, without necessarily advertising this intention.
- In cases of meaning$_L$ (Ladyginian), individuals intentionally reveal their intentions. We are arguing here that most great ape gesture has meaning$_L$.
- In cases of meaning$_G$ (Gricean), individuals intentionally reveal a specifically informative intention. Meaning$_G$ is another name for Grice's 'non-natural meaning', or meaning$_{NN}$. It provides the foundation for linguistic semantics.

Conclusion

A crucial takeaway point is that, with respect to the origins of language, the key comparisons to make between humans and other great apes are not in systems of communication as such, but rather in social cognition, and specifically in means of attention manipulation. Communication systems in nonhuman primates share some surface similarities with natural language, but there are also important dissimilarities which collectively constitute strong evidence against evolutionary continuity (see also Scott-Phillips & Heintz, 2023b). Continuity can rather be identified in social cognition, and more specifically in the domain of attention manipulation, with relatively small differences between humans and other great apes. With the key notion of Ladyginian interaction, we hope to have helped understand these small but very consequential differences.

References

Byrne, R. W., Cartmill, E., Genty, E., Graham, K. E., Hobaiter, C., & Tanner, J. (2017). Great ape gestures: Intentional communication with a rich set of innate signals. *Animal Cognition, 20*(4), 755-769.

Crockford, C., Wittig, R. M., & Zuberbühler, K. (2017). Vocalizing in chimpanzees is influenced by social-cognitive processes. *Science Advances, 3*(11), e1701742.

Fröhlich, M., Wittig, R. M., & Pika, S. (2016). 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.

Geurts, B. (2022). Evolutionary pragmatics: From chimp-style communication to human discourse. *Journal of Pragmatics, 200*, 24-34.

Gómez, J. C. (1994). Mutual awareness in primate communication: A Gricean approach. In: S. T. Parker, R. W. Mitchell, & M. L. Boccia (Eds.), *Self-awareness in Animals & Humans* (pp. 61-80). CUP.

Graham, K. E., Wilke, C., Lahiff, N. J., & Slocombe, K. E. (2020). Scratching beneath the surface: Intentionality in great ape signal production. *Philosophical Transactions of the Royal Society B, 375*(1789), 20180403.

Grice, H. P. (1957). Meaning. *The Philosophical Review, 66*(3), 377-388.

Grice, H. P. (1989). *Studies In The Way Of Words*. HUP.

Heesen, R., & Fröhlich, M. (Eds.) (2022). Revisiting the human 'interaction engine': Comparative approaches to social action coordination. *Philosophical Transactions of the Royal Society B, 377*(1859), 20210092.

Heintz, C., & Scott-Phillips, T. (2023). Expression unleashed: The evolutionary & cognitive foundations of human communication. *Behavioral & Brain Sciences, 46*, e1.

Hobaiter, C., & Byrne, R. W. (2014). The meanings of chimpanzee gestures. *Current Biology, 24*(14), 1596-1600.

Ladygina-Kohts N. N., & de Waal, F., B., M. (2002). *Infant Chimpanzee & Human Child: A Classic 1935 Comparative Study of Ape Emotions and Intelligence*. OUP.

Moore, R. (2017). Convergent minds: Ostension, inference and Grice's third clause. *Interface Focus, 7,* 20160107.

Origgi, G., & Sperber, D. (2000). Evolution, communication and the proper function of language. In: P. Carruthers & A. Chamberlain (Eds.), *Evolution & the Human Mind: Language, Modularity & Social Cognition* (pp. 140-169). CUP.

Scott-Phillips, T. (2015). *Speaking Our Minds*. Palgrave Macmillan.

Scott-Phillips, T., & Heintz, C. (2023a). Great ape interaction: Ladyginian but not Gricean. Proceedings of the National Academy of Sciences, 120(42), e2300243120.

Scott-Phillips, T., & Heintz, C. (2023b). Animal communication in linguistic & cognitive perspective. *Annual Review of Linguistics, 9*, 93-111.

Tauzin, T., & Gergely, G. (2018). Communicative mind-reading in preverbal infants. *Scientific Reports, 8*(1), 1-9

Tomasello, M. (2008). *The Origins of Human Communication*. MIT Press.

Tomasello, M., & Call, J. (2019). Thirty years of great ape gestures. *Animal Cognition, 22*(4), 461-469.

Warren, E., & Call, J. (2022). Inferential communication: Bridging the gap between intentional and ostensive communication in non-human primates. *Frontiers in Psychology, 12*, 718251.

Washburn, M. F. (1908). *The Animal Mind: A Text-Book of Comparative Psychology*. Macmillan Company.

The evolution of gender-differentiated kinship terms

Olena Shcherbakova[*1,4], Scott Claessens[2], Maisy Hallam[3], Marc Allassonnière-Tang[4], and Fiona M. Jordan[5]

[*]Corresponding Author: olena_shcherbakova@eva.mpg.de
[1]MPI for Evolutionary Anthropology, Leipzig, Germany
[2]University of Auckland, Auckland, New Zealand
[3]The University of Edinburgh, Edinburgh, UK
[4]National Museum of Natural History, Lab EA UMR 7206, Paris, France
[5]University of Bristol, Bristol, UK

Recent studies testing hypotheses about cultural variables, like marriage organization, residence and descent patterns, shaping kinship terminology demonstrate that kinship terms have evolved independently from these factors (Passmore & Jordan 2020). So far, alternative explanations of the variation in kinship system have received less attention (but see Kemp et al. 2018).

One alternative explanation involves general linguistic principles. Greenberg (1980) suggests that marked kinship terms cannot have gender distinctions unless they are also present in unmarked kinship terms. Unmarked terms are those that refer to 1) kin closer to the ego (children vs grandparents) and 2) older kin (grandparents vs grandchildren). This leads to the implicational hierarchy: siblings = children > grandparents > grandchildren, meaning that gender distinctions in grandchildren arise after other terms already marking gender. Siblings and children are equally unmarked since the closeness to ego (children) and seniority (siblings) criteria are in conflict. Greenberg (1990: 322) also points out the potential influence of sex-based gender systems on gender-differentiated kinship terms. For instance, the *cousin/cousine* distinction in French might have been facilitated by the presence of a gender system categorizing all nouns into masculine and feminine, which is not the case in English with its gender-neutral *cousin* term.

We test Greenberg's hierarchy of gender-differentiated kinship terms and explore the potential effect of sex-based gender systems on a sample of 303 languages. We map the kinship term data from Kinbank (Passmore et al. 2023) and the sex-based gender data from Grambank (Skirgård et al. 2023) onto the global EDGE tree Bouckaert et al. (2022). We fit Bayesian mixed models using the *brms* package (Bürkner 2017) to establish whether the results support or deviate from Greenberg's predictions.

We find that sex-based gender systems in language are positively correlated only with gender distinctions in grandchildren terms. Greenberg's hierarchy is partially supported: distinctions in grandparents are positively correlated with distinctions in children, but not in siblings, and distinctions in grandchildren are positively correlated with distinctions in grandparents, children, and siblings. This suggests that (1) general linguistic principles proposed by Greenberg can partially account for the variation in the kinship lexicon and (2) gender distinctions in kinship terms (with the exception of children) evolve independently from gender systems.

Acknowledgements

This work was supported by the Department of Linguistic and Cultural Evolution (to O.S.), the French National Research Agency (to M.A-T., grant EVOGRAM: The role of linguistic and non-linguistic factors in the evolution of nominal classification systems, ANR-20-CE27-0021) and the European Research Council (Starting Grant VARIKIN ERC-Stg-639291 to F.M.J.).

References

Bouckaert, R., Redding, D., Sheehan, O., Kyritsis, T., Gray, R., Jones, K. E., & Atkinson, Q. (2022). *Global language diversification is linked to socio-ecology and threat status*. https://osf.io/preprints/socarxiv/f8tr6/

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. Journal of Statistical Software, 80(1), 1–28.

Greenberg, J. (1980). *Language Universals*. Berlin: De Gruyter.

Greenberg, J. (1990). Universals of kinship terminology: Their nature and the problem of their explanation. In K. Denning & S. Klemmer (Eds.), *On language.* (pp. 310–327). De Gruyter.

Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics*, *4*, 109–128.

Passmore, S., Barth, W., Greenhill, S. J., Quinn, K., & others. (2023). Kinbank: A global database of kinship terminology. *Plos One*, *18*(5), e0283218.

Passmore, S., & Jordan, F. M. (2020). No universals in the cultural evolution of kinship terminology. *Evolutionary Human Sciences*, *2*, e42.

Skirgård, H., Haynie, H. J., Blasi, D. E., & others. (2023). Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, *9*(16), eadg6175.

The Development of Trisynaptic Pathway in the Hippocampus and the Milestone of Language Development

Edward Ruoyang Shi[1*], Elizabeth Qing Zhang[2], Lluís Barceló-Coblijn[3] and Rui He[1]

[*]Corresponding Author: edwardshiruoyangend@gmail.com
[1]Department of Translation and Language Sciences, Universitat Pompeu Fabra, Barcelona , Spain
[2]School of Linguistics Sciences and Arts, Jiangsu Normal university, Xuzhou, China
[3]Laboratori d'Investigació en Complexitat i de Lingüística Experimental, Universitat de les Illes Balears, Illes Balears, Palma, Spain

1. Introduction

The hippocampus is involved in both episodic memory (Tulving and Markowitsch, 1998) and language processing (Duff and Brown-Schmidt, 2012). From evolutionary perspective, Zhang and Shi (2021) suggests that the hippocampus serves as the subcortical hub underlying displacement, which makes human language flexible in time and space. However, very few studies have explored the relation between the developmental trajectory of the hippocampus and flexibility of human language. In this study, by reviewing previous researches, we propose that the emergence of trisynaptic pathway in the hippocampus not only forms the basis for episodic memory but also serves as the milestone for the human language flexibility. We further suggest that the protracted development of such trisynaptic pathway serves as a channel for information exchange between episodic memory and language, and further forms the neurological basis for the domain-general mechanism between them.

2. Neurological structure of hippocampus

Within the hippocampus, there are two main pathways that connect the subfields of the hippocampus. The trisynaptic pathway (TSP) connects the entorhinal cortex, dentate gyrus, CA3 and CA1, and the monosynaptic pathway (MSP) connects entorhinal cortex and CA1 (Schapiro et al., 2016). MSP develops earlier than TSP. It has been shown that MSP is mainly related to statistical learning (Ellis et al., 2021) and associative learning which is not continuously maintained in the memory (Gómez and Edgin, 2016). The late emergence of trisynaptic pathway, which occurs around 18-24 month, marks the developmental shift in the functions of the hippocampus to episodic memory. Before this period, even though infants can remember salient event in their lives, but details cannot be maintained (Peterson et al., 2011). While due to the emergence of TSP, learning becomes more flexible and the children can remember an object separate from its learning context (Robinson and Pascalis, 2004).

Beyond the domain of episodic memory. We propose that the emergence of the TSP also marks the milestone of human language flexibility. Studies on language acquisition have shown that during the same period (around 18-24 month) when TSP appears, children show a vocabulary spurt (Goldfield and Reznick, 1990) and grammatical developments (Maez, 1983). We suggest that the developmental transitions in both language and episodic memory is not a coincidence. The parallel of developmental trajectories in both domains implies a deep relation between the two domains subserved by a domain-general mechanism (Zhang and Shi, 2021).

3. Evidence from a Clinical Perspective

The relation between the function of the TSP in the hippocampus and linguistic ability is also implicated in the clinical studies. Dysfunctions in schizophrenic TSP have been reported in previous studies (Benes 1991, 1999). Further, Farmer et al. (2023) explore~~s~~ the ultrastructural organization of the TSP in schizophrenia and found excitatory and inhibitory imbalances in TSP. In Alzheimer’s disease, trisynaptic pathway is also susceptible to premature degeneration (Llorens-Martín et al. 2014). In both schizophrenia and Alzheimer’s disease, in addition to memory impairment, impairment of linguistic ability and reduced linguistic flexibility are also detected (Mckenna and Oh, 2005; Bickel et al., 2000).

4. Conclusion

We hypothesize that the emergence of TSP in the hippocampus gives rise the developmental transition in both memory and language, which captures the flexibility of both domains. This could be the neurological foundation for the generative nature of a domain-general mechanism.

References

Benes FM (1991). Morphometric analyses of the hippocampal formation in schizophrenic brain. *Schizophrenia Bulletin*, 17:597–608

Benes FM (1999). Evidence for altered trisynaptic circuitry in schizophrenic hippocampus. *Biological Psychiatry*, 46(5):589–599

Duff, M.C. and Brown-Schmidt, S., (2012). The hippocampus and the flexible use and processing of language. *Frontiers in Human Neuroscience*, 6, p.69.

Ellis, C.T., Skalaban, L.J., Yates, T.S., Bejjanki, V.R., Córdova, N.I. and Turk-Browne, N.B., (2021). Evidence of hippocampal learning in human infants. *Current Biology*, 31(15), pp.3358-3364.

Farmer, C.B., Roach, E.L., Bice, L.R., Falgout, M.E., Mata, K.G., Roche, J.K. and Roberts, R.C., (2023). Excitatory and inhibitory imbalances in the trisynaptic pathway in the hippocampus in schizophrenia: a postmortem ultrastructural study. *Journal of Neural Transmission*, pp.1-17.

Goldfield, B.A. and Reznick, J.S., (1990). Early lexical acquisition: Rate, content, and the vocabulary spurt. *Journal of Child Language*, 17(1), pp.171-183.

Gómez, R.L. and Edgin, J.O., (2016). The extended trajectory of hippocampal development: Implications for early memory development and disorder. *Developmental Cognitive Neuroscience*, 18, pp.57-69.

Llorens-Martín M, Blazquez-Llorca L, Benavides-Piccione R, Rabano A, Hernandez F, Avila J, DeFelipe J., (2014). Selective alterations of neurons and circuits related to early memory loss in Alzheimer's disease. *Frontiers in Neuroanatomy*. 27;8:38. doi: 10.3389/fnana.2014.00038. PMID: 24904307; PMCID: PMC4034155.

McKenna, P.J. and Oh, T.M., (2005). Schizophrenic speech: *Making sense of bathroots and ponds that fall in doorways*. Cambridge: Cambridge University Press.

Robinson, A.J. and Pascalis, O., (2004). Development of flexible visual recognition memory in human infants. *Developmental Science*. 7, 527-533.

Tulving, E. and Markowitsch, H.J., (1998). Episodic and declarative memory: role of the hippocampus. *Hippocampus*, 8(3), pp.198-204.

Zhang Q, Shi E.R. (2021) Why language survives as the dominant communication tool: A neurocognitive perspective. *Behavioral & Brain Sciences*. 2021 Jan 1;44.

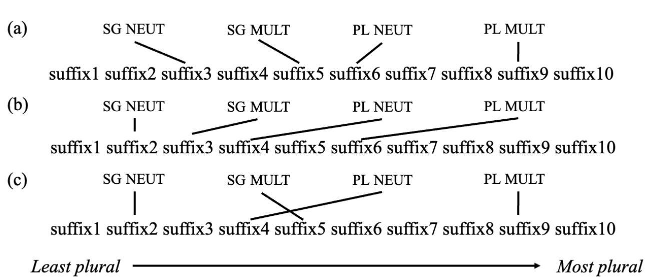
Scalar Morphology: how linguistic complexity can become redundant, yet be actively maintained by analogy

Helen Sims-Williams[*1], Jérémy Pasquereau[1,2] and Matthew Baerman[1]

[*]Corresponding Author: h.sims-williams@surrey.ac.uk
[1]Surrey Morphology Group, University of Surrey, Guildford, United Kingdom
[2]Laboratoire de Linguistique de Nantes, CNRS & University of Nantes, Nantes, France

The Seri language (isolate, Mexico) exhibits a rare feature we call **scalar morphology.** Two independent number features are marked in verbal paradigms (sg/pl subject number and neut/mult event number: Cabredo Hofherr et al 2018; Pasquereau et al 2022), using a common set of around 36 suffixes. At first glance verbal paradigms appear chaotic: depending on the verb, most suffixes can appear in any cell of the paradigm (**disjunctiveness**), and each cell can select almost any suffix (**allomorphy**), making it impossible to associate the suffixes with any consistent element of meaning. However, there is a single way to order the suffixes and the paradigm cells on a scale from 'least plural' to 'most plural' that will yield a monotonic mapping between them (Baerman 2016). This can be visualized as a 'no crossing' constraint: schematically, the mappings represented in Figure 1 (a-b) are possible, but (c) is excluded.

Figure 1. Possible (a-b) and impossible (c) mappings from paradigm cells to suffixes.



The scalar morphology of Seri involves a very high degree of unpredictability of the relationship between linguistic meaning and form, mitigated by a highly unusual constraint. It thus presents a unique challenge for accounts of morphological productivity, learning, and change. Even rare features must arise and persist via general mechanisms of language emergence and change (Newmeyer 2002), and despite its rarity, scalar morphology has been stable across intergenerational transmission in Seri since at least the 1960s (Moser 1961) and is still used productively by speakers to generate and interpret novel forms. Therefore, we ask: (1) How could such a system have come into existence through general mechanisms of language change? (2) Once established, how can it be maintained at a structural level, despite inevitable changes to individual verb paradigms?

We investigated these questions using iterative simulation experiments along the lines of Ackerman & Malouf (2015). Under each model, changes originate as novel predictions for a withheld paradigm cell of a target lexeme, and accumulate over time, potentially leading to structural reorganization. The models differ in the method used for performing these inflectional predictions. In each case, model lexemes are selected. In the **morphemic model**, the target lexeme copies the suffixes used by the model lexemes. Under **set-theoretic analogy**, the target lexeme instead copies an implicational relationship between the paradigm cells of the model lexemes, by completing analogical proportions (Sims-Williams 2022). **Scalar analogy** works in the same way as set-theoretic analogy, except that the paradigm cells are accessed as points on a plurality scale, rather than sets of morphosyntactic features. This licenses a greater range of possible changes than the set-theoretic model, and effectively builds the linear ordering of the paradigm cells into speakers' mental representations. The three models were run on input systems of scalar morphology, and their systemic effects on this input were measured using three evaluation criteria: the proportion of scale violations created, the degree of disjunctive marking, and the degree of allomorphy.

From our results we argue that only an analogical model of morphological productivity and change can account for the persistence of scalar morphology. Moreover, it is capable of doing this without having to 'build in' the linear ordering of cells. Using evidence from internal reconstruction, we suggest that this linear ordering can instead be viewed as a relic of an earlier system in which the forms compositionally marked a single scalar morphosyntactic feature. This feature was reanalysed as two bivalent features (Marlett 2016), and the forms are no longer transparently compositional, but the scale has remained. Seri scalar morphology is thus a prime example of an apparently maladaptive feature of language which is actively maintained by language change, even though its original motivation has long disappeared.

References

Ackerman, F., & Malouf, R. (2015). The No Blur Principle effects as an emergent property of language systems. In *Proceedings of the annual meeting of the Berkeley Linguistics Society* (Vol. 41, No. 41).

Baerman, M. (2016). Seri verb classes: Morphosyntactic motivation and morphological autonomy. *Language*, pp.792-823.

Cabredo Hofherr, P, Pasquereau, J, O'Meara, C. (2018). Event plurality in Seri. In Proceedings of SULA 10, pp. 1-15.

Marlett, Stephen A. (2016). *Cmiique Iitom*: *The Seri language*. Unpublished Ms.

Moser, Edward W. (1961). *Number in Seri verbs*. Philadelphia: University of Pennsylvania master's thesis.

Newmeyer FJ. (2002). Uniformitarian assumptions and language evolution research. In A. Wray (Ed.), *The transition to language*. Oxford: Oxford University Press.

Pasquereau, J, Cabredo Hofherr, P. (2022). Two types of pluractionality in Seri. In *Proceedings of SULA 11.*

Sims-Williams, H. (2022). Token frequency as a determinant of morphological change. *Journal of Linguistics*, 58(3), 571-607.

Simplifications made early in learning can reshape language complexity

Kenny Smith*[1]

*Corresponding Author: kenny.smith@ed.ac.uk
[1]Centre for Language Evolution, The University of Edinburgh, Edinburgh, UK

Languages spoken in larger populations seem to be relatively simple (Wray & Grace, 2007; Trudgill, 2011). One possible explanation is that this is a consequence of the simplifying influence of non-native speakers: adult learners tend to reduce complexity during learning, and large languages tend to have a higher proportion of non-native speakers. This hypothesis, that languages adapt to their social niche (Dale & Lupyan, 2012), receives some statistical support from typological studies which show negative correlations between number of non-native speakers and morphological complexity (e.g. Lupyan & Dale, 2010; Bentz & Winter, 2013; Sinnemäki, 2020, but see ongoing debate in Koplenig, 2019; Kauhanen et al., 2023; Koplenig, 2023). It has also been subjected to experimental tests using artificial language learning techniques, exploring the impact of simplifications made by non-native-like early learners on morphological complexity (Atkinson et al., 2018; Berdicevskis & Semenuks, 2022). Here I report a series of experiments combining their methods, which reconciles the apparent conflict between their results and indicates that the presence of non-native-like early learners in a population can lead to gradual simplification of morphology.

In Experiment 1 I replicate Atkinson et al.'s Experiment 1 using crowdsourced participants (N=94), finding that learners trained on a morphologically complex miniature language simplify its morphology early in learning (e.g. after only 2 bocks of training), but given adequate exposure (e.g. 8 blocks of training) accurately learn the target language. In Experiment 2 I replicate Atkinson et al. Experiment 2: contrary to our original finding, an order of magnitude more data (N=522) suggests that Experiment 2 learners who receive data featuring simplified morphology (produced by early learners from Experiment 1) do show a modest reduction in the morphological complexity of the language they themselves produce. This shows that simplifications made by adult learners can result in simplification of a population's language, although the very small effect size is consistent with Atkinson et al.'s suggestion that learning from the mixed output of multiple individuals reduces or nullifies the simplifications seen in the output of individuals.

However, Atkinson et al. (2018) Experiment 2 (and Experiment 2 here) simulate only a single generation of transmission; other work emphasises the influ-

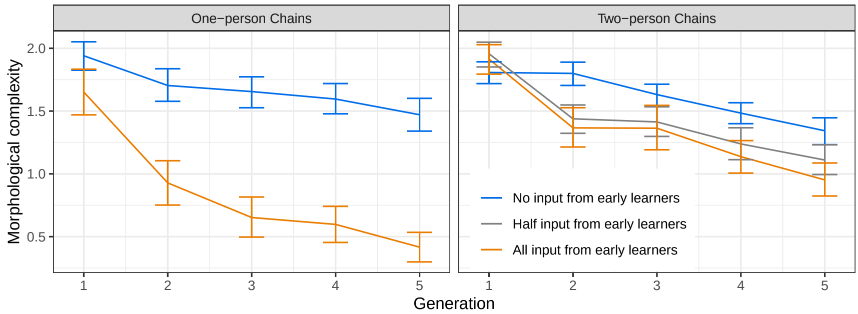


Figure 1. Experiment 3 results, showing morphological complexity (measured using the method from Atkinson et al., 2018) against generation and proportion of input that learners at generation N+1 received from early learners at generation N. Error bars indicate 95% CIs.

ence of multiple generations of transmission in amplifying weak biases in learning (e.g. Reali & Griffiths, 2009). Berdicevskis and Semenuks (2022) therefore use a multi-generation iterated learning paradigm to show that languages transmitted in chains featuring early learners simplify more rapidly than languages from chains featuring no early learners. This result provides support for the hypothesis that simplifications by adult learners could ultimately lead to language simplification. However, Berdicevskis and Semenuks only run chains where each learner is exposed to the output of a single individual at the previous generation, meaning that the mixing mechanism identified in Atkinson et al. is not at play. In Experiment 3 I therefore combine the Atkinson et al. learning paradigm with an iterated learning procedure similar to that used by Berdicevskis and Semenuks, running 5-generation iterated learning chains manipulating (1) the proportion of simplified input each learner receives (i.e. coming from early learners at the previous generation) and (2) the number of individuals in each generation, running both one-person versus two-person chains (N=400, 50 chains). Experiment 3 therefore manipulates proportion of simplified input and allows for mixing effects, as per Atkinson et al., but allows for cumulative effects as per Berdicevskis and Semenuks. Experiment 3 languages gradually simplify (see Figure 1), with simplification being more rapid when learners receive at least some input from early learners at the previous generation. This effect is seen in both one-person and two-person chains, but there is some evidence for mixing effects, with simplification being slower when each learner receives input from multiple individuals.

These results therefore reconcile the apparent mismatch in the experimental literature, being consistent with both Atkinson et al. (2018) and Berdicevskis and Semenuks (2022), and strengthen the experimental evidence for simplification during adult learning as a mechanism which could account for negative correlations between adult learners and linguistic complexity in natural languages.

Acknowledgements

This research received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (Grant 681942).

References

Atkinson, M., Smith, K., & Kirby, S. (2018). Adult learning and language simplification. *Cognitive Science*, *42*(8), 2818–2854.

Bentz, C., & Winter, B. (2013). Languages with More Second Language Learners Tend to Lose Nominal Case. *Language Dynamics and Change*, *3*, 1–27.

Berdicevskis, A., & Semenuks, A. (2022). Imperfect language learning reduces morphological overspecification: Experimental evidence. *PLOS ONE*, *17*(1), e0262876.

Dale, R., & Lupyan, G. (2012). Understanding the origins of morphological diversity: The linguistic niche hypothesis. *Advances in Complex Systems*, *15*, 1150017.

Kauhanen, H., Einhaus, S., & Walkden, G. (2023). Language structure is influenced by the proportion of non-native speakers: A reply to Koplenig (2019). *Journal of Language Evolution*, *8*(1), 90–101.

Koplenig, A. (2019). Language structure is influenced by the number of speakers but seemingly not by the proportion of non-native speakers. *Royal Society Open Science*, *6*(2), 181274.

Koplenig, A. (2023). Still no evidence for an effect of the proportion of non-native speakers on language complexity – A response to Kauhanen, Einhaus & Walkden (2023).

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PLOS ONE*, *5*, e8559.

Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, *111*, 317–328.

Sinnemäki, K. (2020). Linguistic system and sociolinguistic environment as competing factors in linguistic variation: A typological approach. *Journal of Historical Sociolinguistics*, *6*(2), 20191010.

Smith, K., Perfors, A., Fehér, O., Samara, A., Swoboda, K., & Wonnacott, E. (2017). Language learning, language use, and the evolution of linguistic variation. *Philosophical Transactions of the Royal Society B*, *372*, 20160051.

Trudgill, P. (2011). *Sociolinguistic Typology: Social Determinants of Linguistic Complexity.* Oxford University Press.

Wray, A., & Grace, G. W. (2007). The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua*, *117*, 543–578.

Exploring the content of casual Polish conversations

Szala, A.[*1], Wacewicz, S.[1], Placiński, M.[1], Poniewierska, A. E.[1], Schmeichel, A.[1], Stefańczyk, M.[2], & Żywiczyński, P.[1]

[*] Corresponding Author: szala@oakland.edu
[1] Centre of Language Evolution Studies, Nicolaus Copernicus University, Toruń, Poland
[2] Institute of Psychology, University of Wrocław, Poland

In a seminal study, Dunbar, Marriott, and Duncan (1997) found that ≈67% of conversational time is spent on discussing social topics, an estimate that exerted a significant influence on theories of the evolution of the human brain, cognition, and language. However, the work by Dunbar et al. (1997) had substantial limitations (small, demographically and geographically limited sample; data collected exclusively in open environments; unclear operationalisation of "social topics"), which motivated our recent preliminary study on this topic (Szala et al. 2022). Here, we report a follow-up, full-scale study revisiting this issue.

Similarly to Szala et al. (2022), we used Spokes (Pęzik, 2012; 2014). Spokes is a corpus of 669 Polish informal conversations based on live recordings of casual speech obtained in private as well as public places, with speakers from a variety of Polish demographic backgrounds, including age (ranging from 1 to 99 years), and education levels spanning from none to higher education. Some conversations in Spokes were recorded surreptitiously, with consent and demographic data provided after the recording. Spokes is manually divided into lines so as to mark alternating contributions of individual speakers.

In our study, we excluded any conversations too short or too long to reliably code for conversation topics, operationalised as conversations comprising 50 or fewer lines, or containing at least one line with more than 150 word tokens. In the resulting dataset of 535 Polish conversations, every line was coded by two native Polish-speaking coders to ensure reliability. Our main distinction was between social vs non-social topics, understood in terms of information content. Social topics were operationalised as "sharing information related to self and other people", as opposed to sharing other information (e.g., factual conversation about technology). We excluded all lines of text with rater disagreement, i.e. coded as social by one rater but non-social by the other. Two coders agreed on 71% of lines in the dataset. This resulted in a dataset comprising 197,621 lines of text, with a mean conversation length of ca. 367 lines, and a mean line length of ca. 9 word tokens. Our study was pre-registered (Szala et al., 2023; https://osf.io/kjf4e), and the dataset and coding scheme were made publicly available (https://osf.io/mqs5k/).

In contrast to Dunbar et al. (1997), who found that ≈67% of conversation time is spent on social topics, our study indicates that social topics can account for as much as 85% of conversation. An important aspect of our work is how “social topics” are operationalized. In Szala et al. (2022), we *excluded talking about oneself, i.e. counted it as non–social*, and found that 51% of conversations were devoted to social topics so defined. However, since self-disclosure is pivotal in forming social relationships and in particular plays a vital role in reputation building, here we decided to include this category under the rubric “social”. This further underscores the important point that the proportion of "social" to "non-social" topics in conversation is highly sensitive to how "social topics" are defined. In this study, in line with our evolutionarily motivated research question, we defined "social topics" focusing on subjects related to social bonding, cooperation, and human interactions, which may not be universally applicable to all theoretical perspectives on language use.

Our main result (85% of conversation devoted to social topics) confirms, and actually exceeds, Dunbar et al.’s original estimate of 67%, and aligns with other studies examining social discourse, such as 76% in Dahmardeh & Dunbar (2017), reinforcing the general conclusion that a majority of topics in casual conversations tend to be of a social nature. The consistent findings across a range of studies that use varied datasets and methodologies underscore the crucial role that exchanging social information plays in human communication across different contexts and populations. This in turn lends indirect support to theories that stress the role of social ecology in shaping hominin cognitive evolution. We complement our study with exploratory analyses that include demographic factors (gender, age, and education) and a finer subcategorisation of social topics into information on individuals (i) participating in the conversation, (ii) not participating in the conversation but known to the conversants, (iii) not known to the conversants (e.g. celebrities, fictional characters).

Acknowledgements
This research was supported by the Polish National Science Centre under grant agreement UMO-2019/34/E/HS2/00248.

References

Dahmardeh, M. & Dunbar, R. I. M. (2017). What shall we talk about in Farsi? *Human Nature, 28*(4), 423–433. https://doi.org/10.1007/s12110-017-9300-4

Dunbar, R. I. M., Marriott, A., & Duncan, N. D. C. (1997). Human conversational behavior. *Human Nature, 8*(3), 231–246. https://doi.org/10.1007/BF02912493

Pęzik, P. (2012). Jezyk mówiony w NKJP [Spoken language in National Corpus of Polish Language]. In: A. Przepiórkowski, M. Bańko, R. Górski, & B. Lewandowska-Tomaszczyk (Eds.). *Narodowy Korpus Jezyka Polskiego* [National Corpus of Polish Language], 37–47. Wydawnictwo Naukowe PWN, Warszawa.

Pęzik, P. (2014, October 24–25). *Spokes – a search and exploration service for conversational corpus data.* [Selected Papers]. CLARIN 2014: Soesterberg, The Netherlands. www.ep.liu.se/ecp/116/009/ecp15116009.pdf

Szala, A., Placiński, M., Poniewierska, A., Szczepańska, A., & Wacewicz, S. (2022). How much language use is actually on social topics? In: A. Ravignani, R. Asano, D. Valente, F. Ferretti, S. Hartmann, M. Hayashi, … S. Wacewicz, (Eds.). *The Evolution of 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., Placiński, M., Żywiczyński, P., Poniewierska, A. E., Schmeichel, A., & Wacewicz, S. (2023, January 25). *How much language use is actually on social topics: Human conversational behavior revisited.* https://doi.org/10.17605/OSF.IO/KJF4E

Communicative efficiency is present in young children and becomes more adult-like with age

Shira Tal[*1], Kenny Smith[1], Inbal Arnon[2] and Jennifer Culbertson[1]

[*]Corresponding Author: stal@ed.ac.uk
[1] Centre for Language Evolution, University of Edinburgh, UK
[2] Psychology Department, The Hebrew University of Jerusalem, Israel

One of the most striking commonalities across the world's languages is the tendency to assign less linguistic material to more predictable or frequent meanings (Gibson et al., 2019; Haspelmath, 2021). The association between form length and meaning is argued to derive from speakers' bias for efficient communication, reflecting the need to balance competing pressures: minimizing production effort while maximizing understandability (Levshina & Moran, 2021; Zipf, 1949). An efficient trade-off between these two pressures involves producing less linguistic material whenever possible, e.g., when the meaning is predictable, and producing more linguistic material only when it is essential for being understood, e.g., when the meaning is unpredictable or when there is noise. Indeed, there is abundant evidence showing that speakers tend to reduce or omit elements when that does not compromise understandability, and tend to use longer forms when shortening or omitting them would impede understanding (Kanwal et al., 2017; Levshina & Moran, 2021). However, virtually all of this evidence comes from adults. It is not clear whether children's language use is also shaped by communicative efficiency. Investigating whether such a pressure is already present in children is important for understanding both the development of communicative behaviour and the respective roles of adults and children in shaping language structure. Here, we investigate the developmental trajectory of communicative efficiency using a novel experimental paradigm. Children between the ages of 4 and 10 play a communication game with a simulated interlocutor using visual icons: they have to tell the simulated interlocutor which action they should perform when meeting another character (kiss or hit). The design simulates effort and understandability in the following way. To simulate effort, messages can vary in length (1-3 icons), with longer messages taking more effort to produce than shorter messages. To simulate environmental noise, in some communicative turns messages are corrupted, and longer messages are robust to that corruption while shorter messages are not (see Figure 1). If efficient communication is present already in younger children, then longer messages should be used in the presence of noise and shorter messages in non-noisy environments, regardless of age. If, however, this tendency is tied to development, then the relation between noisiness and message length should vary with age. Importantly, communicative efficiency could develop with age in two ways. If young children show a weaker preference to maximize understandability, we should see a developmental increase in the tendency to use longer messages in noisy environments; If young children show a weaker preference to minimize effort, we should see a developmental increase in the tendency to use shorter messages in non-noisy environments. 61 Hebrew-speaking children participated in the experiment (mean age: 6;10y). Results show that communicative efficiency is attested already in young children and becomes more adult-like with age: Even young children produce longer messages in noisier environments, but as they grow, they are more likely to shorten messages (minimize effort) when a short message is sufficient for accurate communication (see Figure 2). We discuss the implications of our results for theories of language evolution and change as well as cognitive development.

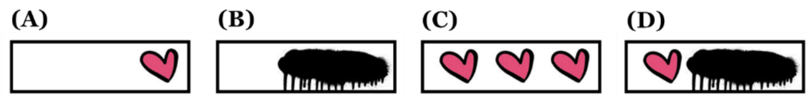


Figure 1: Example message types conveying the kiss action. Note that the noise obscures the first two icons when being "read" from right to left, as in Hebrew (participants' native language). (A) A 1-icon length message, without noise. (B) A 1- or 2-icon length message, with noise. (C) A 3-icon length message without noise. (D) A 3-icon length message with noise.

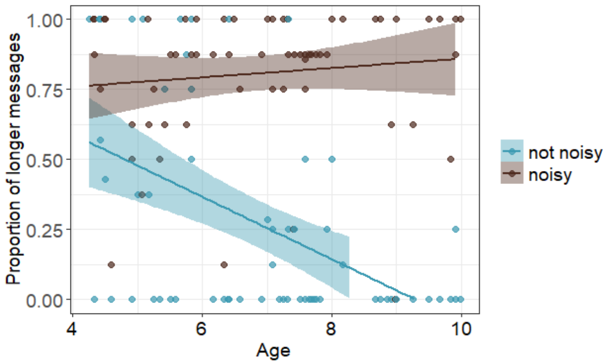


Figure 2: Proportion of longer messages as function of age (in years) and noise. Individual points represent by-participant means. Solid lines show estimated regression lines for noisy and not noisy trials, along with 95% confidence intervals.

References

Gibson, E., Futrell, R., Piandadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How Efficiency Shapes Human Language. *Trends in Cognitive Sciences*, *23*(5), 389–407. https://doi.org/10.1016/j.tics.2019.02.003

Haspelmath, M. (2021). Explaining grammatical coding asymmetries: Form-frequency correspondences and predictability. *Journal of Linguistics*, *57*(3), 605–633. lingbuzz/004531%0A

Kanwal, J., Smith, K., Culbertson, J., & 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.

Levshina, N., & Moran, S. (2021). Efficiency in human languages: corpus evidence for universal principles. *Linguistics Vanguard*, *7*, 1–8.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley Press.

Exploring Systematic Phonological Cues in Language: A Comparative Study Across 60 Languages from 13 Families

Nicole Tamer*[1,2], Paul Widmer[1,2]

*Corresponding Author: nicole.tamer@uzh.ch
[1]Department of Comparative Language Science, University of Zurich, Zurich, Switzerland
[2]Center for the Interdisciplinary Study of Language Evolution, University of Zurich, Zurich, Switzerland

Human language differs from animal communication in many respects, most prominently by having the capacity for great flexibility and arbitrariness in its expression which has evolved in the hominin lineage since speciation from the last common ancestor (Watson et al. 2022; C. F. Hockett and C. D. Hockett 1960). However, at some point in evolutionary history, non-arbitrary constraints have evolved to render language more efficient and easier to process to adapt to the needs of communication over generations by accumulating learning preferences (Kirby, Griffiths, & Smith, 2014; Motamedi et al., 2022).

This differentiation is explored in our study which focuses on the prevalence of systematicity across languages. The research questions are whether systematicity is distinguishing between word classes (open and closed) and if it is language-specific as evidenced by previous studies (Dingemanse et al., 2015). Corpus studies have revealed that some languages showcase systematic constraints, such as subtle systematic phonological cues to differentiate between word classes and phonological categories (Kelly 1992; Monaghan, Chater, et al. 2005). These cues provide cognitive advantages resulting in ease of processing, improved language comprehension and acquisition of languages (Raviv, Heer Kloots, et al. 2021; Fitneva et al. 2009; Monaghan, Christiansen, and Fitneva 2011). Understanding the diversity of systematicity is crucial in uncovering its cognitive advantages, such as enhanced memory processing, learnability, and acquisition (Raviv et al., 2021; Monaghan et al., 2012; Fitneva et al., 2009), as well as its significant role in the emergence and evolution of expansive lexical and grammatical inventories.

Recent studies using novel computational and statistical methods have underscored the increasing relevance of systematicity (Raviv & Arnon, 2018;

Pimentel et al., 2019; Nölle et al., 2018). However, many prior investigations were limited to a narrow sample predominantly biased towards modern Western European languages or analyzed only a limited number of words, constraining the generalizability of findings to other linguistic contexts. To address this gap, we conducted an extensive analysis encompassing grammatical data from 40 modern and ancient Indo-European languages, alongside 20 languages belonging to 12 distinct language families. The data was compiled from language-specific corpora, grammars as well as comparative language data bases. Specifically, we scrutinized phonological cues pertaining in the initial phoneme, thereby capturing the initial word recognition advantages conferred by systematicity (Trott et al., 2019; Tamariz, 2008).
With a Bayesian logistic regression model, we investigated the relationship between phonological cues and systematicity. A strong amount of systematicity is defined as the data points aggregating in the upper quantile of either open or closed class. The posterior probability results show how much evidence there is for a cue within a language occurring either above or below zero. Phonological cues with posterior probability values close to 1.0 are well supported (Greenhill, Gray, et al. 2009). Fig. 1 demonstrates the ubiquity of systematic patterns across all languages, on the clade level and across phonological categories.

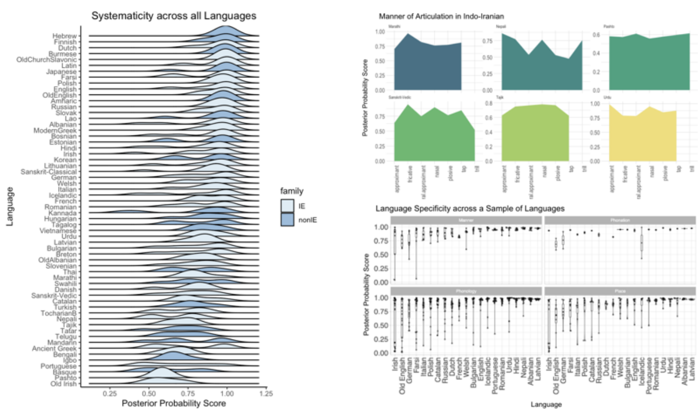


Figure 1. Systematicity patterns across all languages within the Indo-Iranian language clade and across all phonological categories (Place and Manner of Articulation, Phonology and Phonation).

This recurrent pattern was observed in other clades and phonological categories, such as place of articulation, phonation, and individual phonemic units. Consequently, it can be assumed that systematicity is present across all observed languages which seems to be a cross-linguistic pattern of distinct phoneme distribution in initial word segments between open and closed word classes. This effect, however, was not observed within the open word class, contrary to prior research. These discoveries underscore the pervasiveness and diversity of non-arbitrariness in a variety of global languages.

Acknowledgments

This work was supported by the NCCR Evolving Language. Swiss National Science Foundation Agreement Nr. 51N40_180888.

References

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, iconicity, and systematicity in language. *Trends in cognitive sciences*, *19*(10), 603-615.

Fitneva, S. A., Christiansen, M. H., & Monaghan, P. (2009). From sound to syntax: Phonological constraints on children's lexical categorization of new words. *Journal of child language*, *36*(5), 967-997.

Gray, R. D., Drummond, A. J., & Greenhill, S. J. (2009). Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *science*, *323*(5913), 479-483.

Hockett, C. F., & Hockett, C. D. (1960). The origin of speech. *Scientific American*, *203*(3), 88-97.

Kelly, P. (1992). Does the ear assist the eye in the long-term retention of lexis?. *IRAL: International Review of Applied Linguistics in Language Teaching*, *30*(2), 137.

Monaghan, P., Mattock, K., & Walker, P. (2012). The role of sound symbolism in language learning. *Journal of experimental psychology: Learning, memory, and cognition*, *38*(5), 1152.

Monaghan, P., Christiansen, M. H., & Fitneva, S. A. (2011). The arbitrariness of the sign: learning advantages from the structure of the vocabulary. *Journal of Experimental Psychology: General*, *140*(3), 325.

Monaghan, P., Christiansen, M. H., & Chater, N. (2007). The phonological-distributional coherence hypothesis: Cross-linguistic evidence in language acquisition. *Cognitive psychology*, *55*(4), 259-305.

Monaghan, P., Chater, N., & Christiansen, M. H. (2005). The differential role of phonological and distributional cues in grammatical categorisation. *Cognition*, *96*(2), 143-182.

Nölle, J., Staib, M., Fusaroli, R., & Tylén, K. (2018). The emergence of systematicity: How environmental and communicative factors shape a novel communication system. *Cognition*, *181*, 93-104.

Pimentel, T., McCarthy, A. D., Blasi, D. E., Roark, B., & Cotterell, R. (2019). Meaning to form: Measuring systematicity as information. *arXiv preprint arXiv:1906.05906*.

Raviv, L., de Heer Kloots, M., & 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., & Arnon, I. (2018). Systematicity, but not compositionality: Examining the emergence of linguistic structure in children and adults using iterated learning. *Cognition*, *181*, 160-173.

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

Trott, S., Semenuks, A., & Bergen, B. (2019). Sub-morphemic form-meaning systematicity: the impact of onset phones on word concreteness. In *CogSci* (p. 3373).

Watson, S. K., Filippi, P., Gasparri, L., Falk, N., Tamer, N., Widmer, P., ... & Glock, H. J. (2022). Optionality in animal communication: a novel framework for examining the evolution of arbitrariness. *Biological Reviews*, *97*(6), 2057-2075.

Gesture as precursor to language: Evidence from bimodal bilingualism

Seyyed Hatam Tamimi Sa'd [*1,] and Ronnie B. Wilbur[2]

[*]Corresponding Author: shtamimi90@gmail.com
[1&2] Department of Linguistics, Purdue University, West Lafayette, USA

Two hypotheses frequently cited within the frameworks of language origin are (i) vocal origins and (ii) gestural origins (van Schaik, 2016). The vocal hypothesis argues that language arose among hominins as a result of vocal learning after hominins acquired the ability to learn vocalizations socially due to changes that occurred in ape vocalizations. The gestural hypothesis states the features of language evolved among apes in gestural communication and were later transferred to the vocal domain. Supporters of the gestural hypothesis point to the universal occurrence of co-speech gestures among humans (e.g., Corballis, 2013) and shared properties of human language and gestural communication in nonhuman primates (Meguerditchian, Cochet & Vauclair, 2011) as evidence. They further mention the advantages of speech over gesture for the universal presence of language in the auditory-oral modality, e.g., the ability to communicate in the dark or while making tools (Corballis, 2002). If we assume that language had gestural origins and that gesture and speech have co-existed for an extended period, it sounds plausible to predict that the human brain has then evolved to be capable of simultaneously not only producing speech and sign but also of doing so in two different (i.e., mismatching) orders. Investigating this claim requires that the language user have the extraordinary physical ability to produce a proposition via two output channels simultaneously in two typologically different languages. Such an ability is only available to bimodal bilinguals, i.e., individuals competent in a spoken language and a signed language and can thus use speech and sign to express the same sentence (Donati, 2021). In this study, we examined whether the human brain can express the same sentence simultaneously not only in two different modalities but also two different word orders. The two languages are Khuzestani Arabic (KhA) and Sadat Tawaher Sign Language (STSL). STSL is a young sign language that emerged naturally approximately sixty years ago in a family in a small village named *Sadat Tawaher* in southwestern Iran after a man lost his hearing and has since been in use. STSL arose in complete isolation from any other sign language(s) and with none of its signers having had any prior knowledge of or exposure to any sign language. Since STSL signers speak when they sign, we examined whether they would produce their output following the same order or different word orders. The participants were 9 native signers (5 males and 4 females; aged 19-54) who are all hearing and who had learned KhA as their mother tongue. However, they used STSL to communicate with the deaf person who himself did not sign but could speak normally to communicate with others as he lost his hearing long after he had learned his mother tongue, KhA, in his early twenties. This created a unique situation which, to the best of our knowledge, is unattested in the literature. In terms of proficiency, the participants were fully balanced and fluent in both KhA and STSL. Furthermore, comparing KhA and STSL is justified as the two languages are typologically different in that while KhA is head-initial (as reflected in its SVO order, preverbal position of negators, and wh-movement), STSL is head-final (as seen in its sentence-final placement of completive aspect markers, negators, and *wh*-signs). The data consisted of sentence productions and storytelling, totaling 1988 sentences (524 declaratives, 283 *wh*-questions, and 1181 negative sentences). The results showed that, overall, 442 (22%) of the speech/sign strings had a matching order (i.e., congruent lexicalizations) while 1546 strings (78%) had mismatching orders (i.e., incongruent lexicalizations). Furthermore, the signers adhered to the same grammar in 332 declaratives (63%) and 110 *wh*-questions (39%), but used a different word order for all the negative sentences (i.e., 100% mismatch). Our findings clearly indicate that the bimodal bilingual mind not only actively hosts two languages in two different modalities but is also capable of producing them simultaneously in two different orders, via the vocal tract and the hands (and face). The data also indicated cross-modal influence from speech to sign and sign to speech as evidenced in utterances with *wh*-doubling and of bimodal bilinguals' preference for code-blending (i.e., using both speech and sign) over code-switching which is consistent with previous research (e.g., English-ASL, Emmorey, et al., 2003). The results have implications for the architecture of the bilingual mind and what it can achieve. Bimodal bilinguals seem to avoid costly inhibitory control and additional processing costs by taking advantage of the suspension of articulatory constraints. These results might also provide support for gesture as the first phylogenetic precursor of human language and seem to be in line with the findings of some recent studies that have documented the existence of a bimodal intentional communicative system in captive chimpanzees (Hopkins, Taglialatela, & Leavens, 2007). This bimodal intentional system was preceded by communicative gestures in the common ancestor of humans and chimpanzees (Meguerditchian, et al., 2011). In conclusion, the findings provide support for the historical connections between gesture and speech and for a single integrated communication system that oversees both vocal and gestural communication. Whether these findings imply that speech has supplanted gesture over time remains open-ended (see McNeill, 2012).

References

Baker, A., & van den Bogaerde, B. (2008). Codemixing in signs and words in input to and output from children. In C. Plaza Pust & E. Morales López (Eds.), *Sign bilingualism* (pp. 1-27). Amsterdam: John Benjamins.

Corballis, M. C. (2002). *From hand to mouth: The origins of language*. Cambridge, MA: Harvard University Press.

Corballis, M. C. (2013). Gestural theory of the origins of language. In C. Lefebvre, B. Comrie, & H. Cohen (Eds.), *New perspectives on the origins of language* (pp. 171-184). Amsterdam / Philadelphia: John Benjamins Publishing Company.

Donati, C. (2021). Bimodal bilingual grammars: Theoretical and experimental perspectives. In J. Quer, R. Pfau., & A. Herrmann (Eds.), *The Routledge handbook of theoretical and experimental sign language research* (pp. 614-635). Routledge: London & New York.

Emmorey, K., Grabowski, T., McCullough, S., Damasio, H., Ponto, L., Hichwa, R., & Bellugi, U. (2003). Neural systems underlying lexical retrieval for sign language. *Neuropsychologia*, *41*, 85-95.

Hopkins, W. D., Taglialatela, J. P., & Leavens, D. A. (2007). Chimpanzees differentially produce novel vocalizations to capture the attention of a human. *Animal Behaviour*, *73*, 281-286.

McNeill, D. (2012). *How language began: Gesture and speech in human evolution*. Cambridge: Cambridge University Press.

Meguerditchian, A., Cochet, H., & Vauclair, J. (2011). From gesture to language. In A. Vilain, J. Schwartz, C. Abry, & J. Vauclair (Eds.), *Primate communication and human language* (pp. 91-119). Amsterdam / Philadelphia: John Benjamins Publishing Company.

van Schaik, C. P. (2016). *The primate origins of human nature.* NJ: John Wiley & Sons, Inc.

Semantics-based spontaneous compounding emergence in artificial sign languages

Oksana Tkachman*[1], and Carla L. Hudson Kam[1],

*Corresponding Author: o.tkachman@alumni.ubc.ca
[1]Department of Linguistics, The University of British Columbia, Vancouver, Canada

Compounding is one of the simplest and most widespread word-formation processes in human languages: new words are created by combining existing ones in a simple hierarchical structure involving a head and a modifier. Some researchers claim compounding to be one of the earliest linguistic processes to emerge in language evolution (Jackendoff 1999, 2002; Heine and Kuteva 2008), but it remains available in older established languages as well. In spoken languages, compounding is nearly universal (Bauer 2017), and so far, has also been attested in all signed languages where word-formation processes have been investigated (Meir, Aronoff, Sandler and Padden 2010; Tkachman and Meir 2018).

Here, we report on the use of compounding as a word formation strategy in a sign creation study. In the study, 50 native speakers of English with no knowledge of any sign language, ages 19-72, were asked to create sign names for 100 objects, appropriate to be used in an artificial sign language. The stimuli consisted of pictures presented to the participants one-by-one on a computer screen in a random order. We were primarily interested in the influence of various kinds of iconicity in signs, but during coding, we noted that sometimes participants reused already created signs in what appeared to be compound names for objects that were semantically or conceptually related to the earlier meaning. For example, a few participants reused the sign they coined for CAT much later in a compound sign for LION (see Fig. 1). Based on this observation, we decided to analyze the data for compounds in a more systematic way. We classified as potential compounds all instances of multi-sign responses that were used as a label for an object, and that did not include poses, hesitations, body shifts, etc. This comprised 2385 responses (out of the total 4975 responses). We excluded all compounds where one of the signs indicated only size or shape, as these sign combinations appear closer to classifier constructions (common in sign languages) than true compounds. (We are analyzing those responses separately.) Additionally, we excluded all compounds were all the signs were unique, that is, we only included compounds were at least one of the signs used have been coined for an earlier referent. Note that the reused sign did not have to be part of a multi-sign response;

it could have been introduced in a one-sign response (though we did not count single signs as part of the compound family). This left 1042 responses which were grouped into *sign families* based on the reused sign: compound responses that included the same sign were considered a sign family. Data were coded by two independent coders.

Results: All 50 participants created at least some multi-sign responses, and 44 out of 50 reused signs in compounds. Overall, we identified 266 sign families (mean=5.54 per participant in those who created them). Of the 2385 multi-sign responses, 1042 were members of a sign family (23.7 responses per participant in those who created them, out of 100 total responses per person). Semantically, sign families could be classified as belonging to one of 35 categories, with the majority belonging to vehicles (30 sign families), water-related entities (25), trees (23), hard/solid entities (21), and animals (19). We speculate that these particular categories all have a *prototypical member* which can be easily used as part of a compound for other members of the category. This is especially interesting, since the order of presentation of pictures for sign creation was random; that is, we did not manipulate the data to have more prototypical members to appear before less prototypical members. Another possibility is that the participants reused signs due to *expanding meaning space*: since they did not know which referents they needed to name in the future, or how many, it encouraged them to re-use signs they already created, and to do so in a more systematic manner (see Nölle et al. 2018; Raviv et al. 2019). These findings suggest that some systematicity can emerge without iterative communication, generational transmission, or new learners (cf. Kirby et al. 2015; Raviv et al. 2019).

The coding and analysis of headedness in this dataset is ongoing.

(a)  (b) 

Figure 1. (a) CAT represented with 'whiskers', (b) LION represented with the form for 'whiskers' previously coined for CAT and another sign for 'mane'. Note that this participant was thinking aloud as she was inventing her signs, and explicitly named signs in (b) as 'cat' and 'mane'.

References

Bauer, L. (2017). *Compounds and compounding*. Vol. 155. Cambridge University Press.

Heine, B., & Kuteva, T. (2008). Constraints on contact-induced linguistic change. *Journal of Language Contact*, *2*(1), 57-90.
Jackendoff, R. (1999). Possible stages in the evolution of the language capacity. *Trends in Cognitive Sciences*, *3*(7), 272-279.
Jackendoff, R. (2002). What's in the lexicon?. In S. Nooteboom, S. G. Nooteboom, F. Weerman, & F. N. K. Wijnen (Eds.). *Storage and computation in the language faculty* (pp. 23-58). Springer Netherlands.
Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87-102.
Klima, E. S., & Bellugi, U. (1979). *The signs of language*. Harvard University Press.
Nölle, J., Staib, M., Fusaroli, R., & Tylén, K. (2018). The emergence of systematicity: How environmental and communicative factors shape a novel communication system. *Cognition*, *181*, 93-104.
Raviv, L., Meyer, A., & Lev-Ari, S. (2019). Compositional structure can emerge without generational transmission. *Cognition*, *182*, 151-164.
Supalla, T. (1986). The classifier system in American Sign Language. In Colette G. Craig (Ed.), *Noun classes and categorization* (pp. 181–214). Amsterdam, NL: John Benjamins.
Tkachman, O., & Meir, I. (2018). "Novel compounding and the emergence of structure in two young sign languages." *Glossa: a journal of general linguistics* 3, no. 1.
Zwitserlood, I. (2012). Classifiers. In R. Pfau, M. Steinbach, and B. Woll, (Eds.), *Sign language: An international handbook* (pp. 158-186). Walter de Gruyter.

Compositionality and Algorithmic Complexity in Neural Networks

Charles Torres*[1] and Richard Futrell[1]

*charlt4@uci.edu
[1]Language Science, University of California, Irvine, USA

1. Introduction

Language is compositional. One explanation for the existence of compositionality is that it is more compressible than a holistic system and has a simpler representation when making use of the *minimum description length* (MDL) principle (Brighton, 2002). The MDL principle is a learning procedure where a learner chooses the hypothesis which best optimizes the compressibility of the data and the hypothesis used to generate that data.

However, the MDL principle is sensitive to the hypothesis space under consideration as the computational formalism used can affect the learning procedure (McGregor, 2014). We investigate the compressibility of compositionality in one learning system: LSTMs. Previous work has shown that LSTMs prefer to learn compositional mappings from meaning to signal (Ren, Guo, Labeau, Cohen, & Kirby, 2019). We introduce a method which allows us to measure the complexity/accuracy tradeoff of LSTMs and use this to measure compositional and holistic mappings. We find that compositional languages have a more favorable tradeoff curve within LSTM architectures which may explain this preference.

2. Methods

Measuring the complexity of an LSTM could take a number of different forms. We consider the number of non-zero parameters in the network which measures the number of connections between neurons. We use a stochastic method of approximating L_0-regularization which pressures a network to zero-out parameters Louizos, Welling, and Kingma (2018). The severity of network pruning depends on the strength of the regularization term. See supplementary materials for implementation.

The task was to correctly learn a compositional or holistic language represented as a meaning-to-word mapping. Each meaning was a pair of the form (x, y) where $x, y \in \{1, 2, 3, 4\}$, and each word was an 8 character string from $\{a, b, c\}$. Compositional languages could be decomposed into two 4 character subwords mapping to meanings in x or y. LSTMs were trained using the cross-entropy objective. The entire language formed both the training and testing set

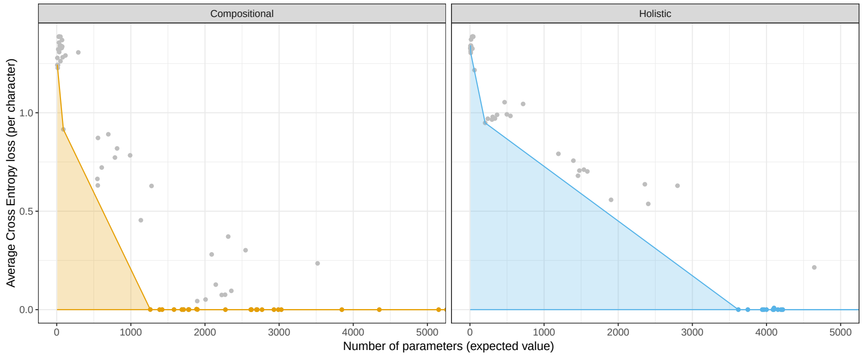


Figure 1. Area under the curve for the holistic and compositional languages. The area under the holistic curve is significantly larger owing to the greater complexity of networks which perform well on the task (colored points).

(memorization being desirable), and training was performed for 20,000 epochs or until performance did not improve for 50 epochs, then accuracy and number of non-zero parameters were stored. Training was repeated with different regularization penalties to yield 80 datapoints per language (for a total of 160). A concave curve was fit to the points representing the frontier of the complexity and accuracy of the LSTMs on the holistic and compositional objectives. The favorability of the tradeoff was measured as a difference in area under these curves.

3. Results

Results are visible in Figure 1. Holistic languages require more neural-network connections to learn accurately when compared with their compositional counterparts. This result is significant under a permutation test ($p < .001$).

4. Conclusion

Compositional languages are easier to represent than holistic languages in a neural network model, a fact which may explain their preferential learning as in Ren et al. (2019). Furthermore, that there is a correspondence between a “neural” representation and a “symbolic” one (as in Brighton (2002)) is interesting. LSTMs are not human brains, but these correspondences may hold in more biologically inspired models (like spiking neural networks), and we believe further investigations into the biases of different formalisms (LSTMs, brains, or otherwise) under simplicity constraints is desirable.

References

Brighton, H. (2002). Compositional syntax from cultural transmission. *Artificial life*, *8*(1), 25–54.

Louizos, C., Welling, M., & Kingma, D. P. (2018). Learning sparse neural networks through l_0-regularization. In *International conference on learning representations.*

McGregor, S. (2014). Natural descriptions and anthropic bias: Extant problems in solomonoff induction. In *Language, life, limits: 10th conference on computability in europe* (pp. 293–302).

Ren, Y., Guo, S., Labeau, M., Cohen, S. B., & Kirby, S. (2019). Compositional languages emerge in a neural iterated learning model. In *International conference on learning representations.*

Gender differences in linguistic complexity through time

Freek Van de Velde[*1]

[*]Corresponding Author: freek.vandevelde@kuleuven.be
[1]Dept. of Linguistics, Research Group QLVL, KU Leuven, Leuven, Belgium

A large swath of linguistic studies have documented gender differences in language use, at the phonological, lexico-grammatical and stylistic level, both in spoken discourse and in written modes. (Labov 1990; Tannen 1994; Rayson et al. 1997; Coates 1998; Biber et al. 1998; Biber & Burges 2000; Härnqvist et al. 2003; Pennebaker et al. 2003; Newman et al. 2008; Yuasa 2010; Keune 2013; Warriner et al. 2013; Podesva & Kajino 2014; Verheijen & Spooren 2017; Hilte et al. 2020, 2022). Recurrent in the studies on the lexico-grammatical level is that men are more likely to display higher complexity, as indicated by average word frequency, type/token ratios, morphological complexity, and syntactic structure. This has been interpreted as men engaging more in 'report (informative) style' and women more in 'rapport (involved) style' (Tannen 1994; Biber and Burges 2000; Brownlow et al. 2003).

What motivates these gender differences? In addition to explanations in terms of gender as a social construct, where men and women conform to implicit or explicit norms and expectations, evolutionary-based accounts have recently been put forth as well in linguistics (Miller 2002; Foolen 2005; Rosenberg and Tunney 2008; Piersoul & Van de Velde 2021). In the latter line of research, language is seen as a costly trait, and verbal display can be used as a reliable fitness cue (in the sense of Zahavi 1997). While women outperform men in linguistic abilities, on average, men are more likely to use language display and women seem to use this cue in mate selection (Miller 2002: Ch. 10; Dunbar et al. 1997; Rosenberg and Tunney 2008; Lange 2011, Lange et al. 2014).

An historical angle on the gender differences in complexity can shed light on how fluid this evolutionary-evolved difference is. If the gender differences evolve in lock-step with the societal developments, for instance showing a convergence between the genders when societal gender roles become less segregated and more malleable, then cultural factors weigh in more heavily.

This study looks at 120 years of written discourse by prolific writers in a 200 million word corpus of Dutch journalistic prose (CCLAMP, Piersoul et al. 2021), assessing aggregate measures of lexical, morphological and grammatical complexity, using the Tscan software (Pander Maat et al. 2014): lexical diversity (adjusted type/token ratio), average word frequency, morphological complexity, the number of abstract and general nouns and verbs, the hierarchical depth of embedding of composite clauses. This aggregate perspective, where we measure different complexity metrics, allows us to extrapolate beyond the findings of earlier diachronic studies into particular constructions (e.g. Palander-Collin 1999).

For each linguistic complexity metric, a linear mixed model was built, with an interaction effect of the year of publication and the gender as the explanatory fixed effect, and a random effect for the individual author and for the journal, to account for personal and editorial style. These models detect (i) a consist effect size and sign on most of the syntactic and lexical measures (i.e. all measures except for word frequency and clause length), with men displaying more complex language, except for lexical diversity, and (ii) a diminishing gender gap, in the course of the 19th and 20th century, no all the metrics except for clause length. For all the lexical measures, women converge to men, in line with fundings by Degaetano-Ortlieb et al. (2021). For the syntactic measures, either both genders approach each other, or men converge to women. The net result is that on all levels that display a statistically robust difference between the two genders in the 19th century, the difference has evaporated in the late 20th century.

These results can be interpreted as showing that the evolutionary account of Miller (2002) is not fully stable across time and culture, though the direction and the effect size of the difference in the 19th century is in conformity with such accounts. We have no ready explanation for the various convergence patterns in the lexicon and in the syntactic domain.

References

Biber D., Conrad, S. & Reppen, R. (1998). *Corpus linguistics: investigating language structure and use*. Cambridge: CUP.

Biber, D. & Burges, J. (2000). Historical change in the language use of women and men: gender differences in dramatic dialogue. *Journal of English Linguistics, 28(1),* 21-37.

Brownlow, S., Rosamond, J. & Parker, J. (2003). Gender-linked linguistic behavior in television interviews. *Sex Roles, 49,* 121-132.

Coates, J. 1998. *Language and gender: a reader*. Oxford: Blackwell.

Degaetano-Ortlieb, S., Säily, T. & Bizzoni, Y. (2021). Registerial adaptation vs. innovation across situational contexts: 18th century women in transition. *Frontiers in Artificial Intelligence, 4,* 609970.

Foolen, A. (2005). Language origins and sexual selection. In: H. Jacobs & L. Wetzels (Eds.), *Liber Amicorum Bernard Bichakjian* (pp. 37-58). Maastricht: Shaker.

Härnqvist, K., Christianson, U., Ridings, D. & Tingsell, J.-G. (2003). Vocabulary in interviews as related to respondent characteristics. *Computers and the Humanities, 37,* 179-204.

Hilte, L., Daelemans, W. & Vandekerckhove, R. (2020). Lexical patterns in adolescents' online writing: the impact of age, gender, and education. *Written Communication, 37(3),* 365-400.

Hilte, L., Vandekerckhove, R. & Daelemans, W. (2022). Linguistic accommodation in teenagers' social media writing: convergence patterns in mixed-gender conversations. *Journal of Quantitative Linguistics, 29(2),* 241-268.

Keune, K. (2013). *Explaining register and sociolinguistic variation in the lexicon: corpus studies on Dutch*. Utrecht: LOT Dissertation.

Labov, W. (1990). The intersection of sex and social class in the course of linguistic change. *Language Variation and Change, 2(2),* 205-254.

Lange, B.P. (2011). Male proneness to verbal display production. *Acta Linguistica, 5,* 97-104.

Lange, B.P., Zaretsky, E., Schwarz, S. & Euler, H.A. (2014). Words won't fail: experimental evidence on the role of verbal proficiency in mate choice. *Journal of Language and Social Psychology, 33(5),* 482-499.

Miller, G. (2002). *The mating mind: How sexual choice shaped the evolution of human nature*. New York: Doubleday.

Newman, M., Groom, C., Handelman, L. & Pennebaker, J. (2008). Gender differences in language use: an analysis of 14,000 text samples. *Discourse Processes, 45,* 211-236.

Palander-Collin, M. 1999. Male and female styles in 17th century correspondence: I THINK. *Language Variation and Change, 11,* 123-141.

Pander Maat, H., Kraft, R., Van den Bosch, A., Van Gompel, M., Kleijn, S., Sanders, T. & Van der Sloot, K. (2014). T-Scan: a new tool for analyzing Dutch text. *Computational Linguistics in the Netherlands Journal, 4,* 53-74.

Pennebaker, J., Mehl, M. & Niederhoffer, K. (2003). Psychological aspects of natural language use: our words, our selves. *Annual Review of Psychology, 54,* 547-577.

Piersoul, J., De Troij, R. & Van de Velde, F. (2021). 150 Years of written Dutch: the construction of the Dutch corpus of contemporary and late modern periodicals. *Nederlandse Taalkunde, 26(3),* 339-362.

Podesva, R. & Kajino, S. 2014. Sociophonetics, gender, and sexuality. In: S. Ehrlich, M. Meyerhoff and J. Holmes (Eds.), *The handbook of language, gender and sexuality*. 2nd edn. (pp. 103-122). Chichester: Wiley Blackwell.

Rayson, P., Leech, G. & Hodges, M. (1997). Social differentiation in the use of English vocabulary: some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics, 2(1),* 133-152.

Rosenberg, J. & Tunney, R.J. (2008). Human vocabulary use as display. *Evolutionary Psychology 6(3):* 538-549.

Tannen, D. 1994. *Gender and discourse*. Oxford: Oxford University Press.

Verheijen, L. & Spooren, W. (2017). The impact of WhatsApp on Dutch youths' school writing. In: E. Stemle and C. Wigham (Eds.), *Proceedings of the 5th conference on CMC and social media corpora for the humanities (cmccorpora17)* (pp. 6-10).

Warriner, A.B., Kuperman, V. & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods, 45,* 1191-1207.

Yuasa, I.P. (2010). Creaky voice: a new feminine voice quality for young urban-oriented upwardly mobile America in women? *American Speech, 85,* 315-337.

Zahavi, A. (1997). *The handicap principle: a missing piece of Darwin's puzzle*. Oxford: Oxford University Press.

Evolving a higher efficiency lexicon: High resource-cost sounds are preferentially allocated to word beginnings and stressed syllables

Andrew Wedel[*1], Adam King[2]

[*]Corresponding Author: wedel@arizona.edu
[1]Linguistics Department, University of Arizona, Tucson, USA
[2]GumGum, Santa Monica, USA

Sound categories differ in the amount of effort required to produce an easily perceptible signal. For example, /k/ is a relatively low resource-cost sound, while /g/ is higher cost because it requires more physiological effort to maintain voicing at the velar position in the vocal tract. This greater cost has been linked to the observation that costly sounds like /g/ are not only more likely to be absent from a language's phoneme inventory, but also less frequently encountered within languages that have them (Everett 2018, 2020). In an information transfer system in which symbols have different costs, communication is most efficient when costly symbols are less frequent overall, while at the same time being preferentially allocated to positions in which they can convey the most information (Zipf 1949, among many others). Relative greater use of high cost sounds in potentially information rich positions results in a more balanced range of sound contrasts at these positions, creating a higher entropy (i.e., more informative) lexicon overall (Shannon 1949). Here we show that higher cost sounds are preferentially allocated to two disparate positions which are high information for entirely different reasons, creating a higher entropy system in positions where it matters most.

(i) Word position: Listeners process sounds incrementally, where each successive sound gives a listener information that helps exclude incompatible words as the speech stream progresses (van Son & Pols 2003; Magnuson et al. 2007). As a consequence of this progressive exclusion, sounds early in a word convey on average more information than those later in the word (King & Wedel 2020).

(ii) Syllable stress: Stressed syllables tend to be more perceptually salient than unstressed syllables, for example by being longer, louder, and/or higher pitched

(Gordon & Roettger 2017). Through these combinations of attributes, stressed syllables provide a superior platform for the perception of phonetic cues, independently of their position in the word.

We assembled a data set of nine genetically diverse phonemically coded languages as a basis for testing the hypotheses that higher-cost sounds are preferentially allocated (i) to word beginnings, and (ii) to stressed syllables. As a proxy for resource-cost, we employed the usage rate of a phoneme within languages, that is, the lexical type frequency of a phoneme (Everett 2018). As predicted, we find a highly significant relationship between usage rate of sounds (on the y-axis) and syllable number (on the x-axis) in this data set, that is, higher cost sounds occur most often in early syllables, all else being equal (Figure 1).

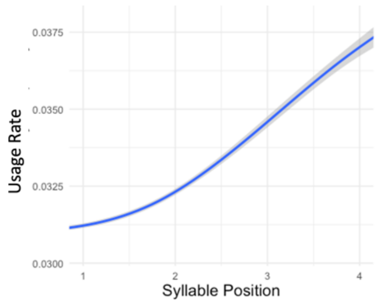


Figure 1. Usage rate by syllable position in the word. Usage rate is lower early in the word indicating greater average cost.

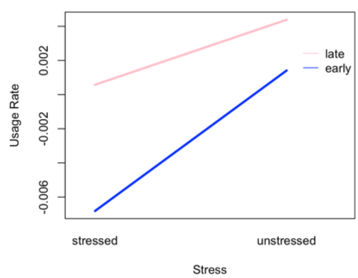


Figure 2. Usage rate by syllable stress level. Stress is associated with lower usage rate, especially at word beginnings.

In addition, we find that stressed syllables are significantly associated with a greater proportion of low usage rate sounds, corresponding to a greater entropy system in stressed positions (Piantadosi et al. 2009). This effect significantly interacts with word position in this dataset: it is strongly negative at the beginnings of words but less so at ends, where the potential to convey information is necessarily lower due to incremental processing (Figure 2; note that usage rate is higher overall late in the word). At the beginnings of words, at which the potential to convey information is high, there are significantly more low usage rate sounds in stressed relative to unstressed syllables. How does this arise? High-information sounds tend to be hyperarticulated, while lower-information sounds tend to be reduced or deleted (e.g., Kanwal et al. 2017, Wedel et al. 2019). Through this process, we expect information-rich positions to more often retain high cost sounds over time, leading over time to a lexicon which is more efficient in its distribution of sounds over words (Dautriche et al. 2017; Gibson et al. 2019).

References

Dautriche, I., Mahowald, K., Gibson, E., Christophe, A., & Piantadosi, S. T. (2017). Words cluster phonetically beyond phonotactic regularities. *Cognition, 163,* 128-145.

Everett, C. (2018). The similar rates of occurrence of consonants across the world's languages: A quantitative analysis of phonetically transcribed word lists. *Language Sciences, 69*, 125-135.

Everett, C. (2018). The global dispreference for posterior voiced obstruents: A quantitative assessment of word-list data. *Language, 94(4),* e311-e323.

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in cognitive sciences, 23(5),* 389-407.

Gordon, M., & Roettger, T. (2017). Acoustic correlates of word stress: A cross-linguistic survey. *Linguistics Vanguard, 3(1),* 20170007.

Kanwal, J., Smith, K., Culbertson, J., & 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.

King, A., & Wedel, A. (2020). Greater early disambiguating information for less-probable words: The lexicon is shaped by incremental processing. *Open Mind, 4*, 1-12.

Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K. & Aslin, R. N. (2007) The dynamics of lexical competition during spoken word recognition. *Cognitive Science 31*, 133-156.

Piantadosi, S. T., Tily, H. J., & Gibson, E. (2009). The communicative lexicon hypothesis. In *The 31st annual meeting of the Cognitive Science Society (CogSci09)*, pp. 2582-2587.

Van Son, R. J. J. H. & Pols, Louis C. W. (2003). How efficient is speech. *Proceedings of the institute of phonetic sciences, 25*, 171-184. University of Amsterdam.

Zipf, G. (1949). *Human behavior and the principle of least effort*. New York: Addison-Wesley

From play pants to pragmatics: Laughter's linguistic leap

Sasha Winkler*[1], Erica Cartmill[1,2], and Greg Bryant[3]

*Corresponding Author: sashawinkler@g.ucla.edu
[1]Department of Anthropology, UCLA, Los Angeles, USA
[2]Department of Anthropology, Cognitive Science Program, Indiana University, Bloomington, USA
[3]Department of Communication, UCLA, Los Angeles, USA

Laughter is a universal human behavior with a unique relationship to language (Bryant & Bainbridge, 2022). It is a pre-linguistic signal, developing in infants before language, and is produced by congenitally deaf people (Provine, 2000). Laughter also has deep evolutionary roots. There is strong acoustic and behavioral evidence that laughter is homologous to the play vocalizations of other great apes. It likely evolved from a play-specific, pant-like signal in our shared primate ancestors (Davila-Ross, Owren, & Zimmermann, 2010). However, in humans, laughter can also function as a pragmatic signal to facilitate turn taking and indicate speaker meaning in language use (Bryant & Bainbridge, 2022). Thus, laughter presents an intriguing case in which a pre-linguistic play vocalization became integrated into the language system.

To better understand this relationship, we conducted a systematic review of play vocalizations across species and the relationships to human laughter (Winkler & Bryant, 2021). Our analysis added support to the theory that human laughter and its precursor in primates could have initially evolved from a cue of heavy breathing during play. We found that acoustic signaling during social play is extensive across the animal kingdom. While animals used a wide variety of sounds to communicate during play, a commonality among many mammals was to use pant-like vocalizations (e.g., short, rhythmic, low-frequency, noisy calls linked to the breath). These vocalizations are used to initiate social play and reduce uncertainty in play interactions in order to avoid aggression. Proximately, play vocalizations appear to be mediated by positive affect (as shown in optimism bias experiments, e.g., Saito, Yuki, Seki, Kagawa, & Okanoya, 2016).

How did a vocalization that is emotionally linked and context-dependent become a flexible feature of human communication? We argue that since diverging from other apes, human laughter has evolved unique functions. It is routinely used outside of social play, within speech, and to broadcast social information. Unlike other types of nonlinguistic emotional vocalizations like crying and screaming, laughter is tightly integrated with speech. It occurs in specific, rule-governed patterns relative to other linguistic constituents to punctuate utterances, signal turn-taking, and indicate irony, humor, and other indirect meanings (Provine, 2000). A likely explanation for this is that humans have different vocal production modes allowing for volitional forms of all nonverbal vocalizations. In the case of laughter, spontaneous and volitional forms are acoustically distinct, perceived differently by listeners, and underpinned by independent neural circuits, with volitional laughter generated by the speech production system (Ackermann, Hage, & Ziegler, 2014; Bryant & Bainbridge, 2022). Volitional laughter became incorporated into our speech production suggesting that its signaling functions are interrelated with functions of language-based conversational interaction more generally.

With greater clarity on laughter's phylogenetic origins comes greater pressure for language evolution theories to account for the puzzle of laughter. Research on laughter—its evolutionary history, development, and neural control—can shed light on the ways that language co-evolved with our ancestors' existing repertoires of vocal communication. In turn, research on the origins of language may help clarify the routes by which laughter came to have pragmatic functions.

References

Ackermann, H., Hage, S. R., & Ziegler, W. (2014). Brain mechanisms of acoustic communication in humans and nonhuman primates: An evolutionary perspective. *Behavioral and Brain Sciences, 37*(6), 529–546.

Bryant, G. A., & Bainbridge, C. M. (2022). Laughter and culture. *Philosophical Transactions of the Royal Society B, 377*(1863), 20210179.

Davila-Ross, M., Owren, M. J., & Zimmermann, E. (2010). The evolution of laughter in great apes and humans. *Communicative & Integrative Biology, 3*(2), 191–194.

Provine, R. R. (2000). *Laughter: A scientific investigation*. NY: Penguin Books.

Saito, Y., Yuki, S., Seki, Y., Kagawa, H., & Okanoya, K. (2016). Cognitive bias in rats evoked by ultrasonic vocalizations suggests emotional contagion. *Behavioural Processes, 132*, 5–11.

Winkler, S. L., & Bryant, G. A. (2021). Play vocalisations and human laughter: a comparative review. *Bioacoustics*, *30*(5), 499-526.

Word-like units emerge through iterated sequence learning

Lucie Wolters[*1], Simon Kirby[2], and Inbal Arnon[1]

[*]Corresponding Author: lucia.wolters@mail.huji.ac.il
[1]Department of Psychology, The Hebrew University, Jerusalem, Israel
[2] School of Philosophy, Psychology and Language Sciences, The University of Edinburgh, Edinburgh, UK

Language exhibits systematicity in the way sets of discrete units, such as words or syllables, are used and reused to form linguistic sequences. This kind of systematicity is reflected in the distributional statistics of language, which in turn provides cues that can help learners discover the building blocks of language – a crucial milestone in language learning. Previous work has shown that a difference in within-unit and between-unit transitional probabilities (e.g. Fiser & Aslin, 2002; Saffran et al., 1996) and a skewed frequency distributions of these units (e.g. Lavi-Rotbain & Arnon, 2021, 2022) can facilitate segmentation in both linguistic and non-linguistic learning domains. However, it is still unexplained how linguistic units and their distributional properties arise in language in the first place. Here, we investigate experimentally whether their emergence may be driven by domain-general constraints on sequence learning over the course of language being repeatedly learned and transmitted.

We conducted an online non-linguistic iterated sequence learning experiment based on Cornish, Smith, and Kirby (2013) in which participants observed and reproduced sets of color sequences that were produced by a previous participant. In each trial, a participant was shown a sequence, made up of four possible colors (red, yellow, green, and blue), and asked to immediately reproduce it. Each participant reproduced a set of 30 sequences, which was transmitted to the next participant. We collected data for 10 transmission chains of 10 generations. The sequences in the initial sets had a length of 12 and were randomly generated.

As is typical in iterated learning experiments, we found a decrease in reproduction error over generations (see figure 1), indicating that the sets of sequences evolved to become easier to reproduce. This was found after accounting for sequence length. To extract units (sub-sequences) from the sequence sets, we used a segmentation method developed by Arnon and Kirby (2024) that segments individual sequences based on the transitional probabilities of the colors the sequence sets. Unit boundaries were created when there was a drop in probability (see figure 1) – similar to how word-boundaries in natural language are often found where the probability of syllable transitions are low (shown in e.g. Stärk et al., 2022). We found that statistically coherent units emerge over iterations and that the distribution of these units became increasingly more skewed, reflected by a decrease in unit entropy (see figure 1). Moreover, the distribution of units showed an increasingly better fit to a Zipfian distribution over time, the typical distribution of word frequencies across languages (Mehri & Jamaati, 2017). Importantly, unit entropy was highly correlated with reproduction error, indicating that the distributional structure of the sequence sets increased their learnability. In addition to these results, I will explore different methods to extract units from sequences and in doing so, contrast the outcomes of transition- and chunking-based learning strategies.

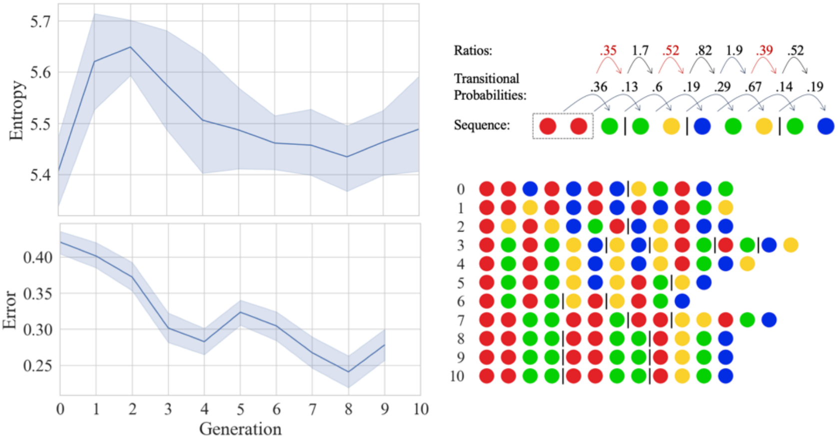


Figure 1. Left-top: Drop in mean unit entropy of chains over generations. Left-bottom: Drop in mean transmission error of chains over generations Right-top: Visualization of the segmentation method used to segment the sequences into units. Right-bottom: The change of a single sequence (out of a set of 30), over the 10 generations. The figure shows sequence 5 from chain D.

Taken together, our findings suggest that domain-general learning pressures during cultural transmission can shape the distributional structure of language and with that, that aspects of linguistic structure may emerge independently of the structure in the meanings that are being conveyed. These results lead to interesting predictions on the emergence of the distributional structure in other culturally transmitted behaviors, such as music.

References

Arnon, I., & Kirby, S. (2024). Cultural evolution creates the statistical structure of language. Scientific reports, 14(1), 5255. https://doi.org/10.1038/s41598-024-56152-9

Cornish, H., Smith, K., & 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, 35(35).

Fiser, J., & Aslin, R. N. (2002). Statistical learning of higher-order temporal structure from visual shape sequences. Journal of Experimental Psychology: Learning, Memory, and Cognition, 28(3), 458–467.

Lavi-Rotbain, O., & Arnon, I. (2021). Visual statistical learning is facilitated in Zipfian distributions. Cognition, 206, 104492.

Lavi-Rotbain, O., & Arnon, I. (2022). The learnability consequences of Zipfian distributions in language. Cognition, 223.

Mehri, A., & Jamaati, M. (2017). Variation of Zipf's exponent in one hundred live languages: A study of the Holy Bible translations. Physics Letters A, 381(31), 2470–2477.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. Science, 274(5294), 1926–1928.

Stärk, K., Kidd, E., & Frost, R. L. A. (2022). Word Segmentation Cues in German Child-Directed Speech: A Corpus Analysis. Language and Speech, 65(1), 3–27.

The Evolution of Language and Human Rationality

Robert Worden

Wellcome Centre for Human Neuroimaging, Institute of Neurology, University College London, London, United Kingdom

Language evolved not just by natural selection, but partly by sexual selection for the display of superior intelligence. This accounts for the uniqueness of human language: other species have not faced the same sexual selection pressures. If language is to be used to display intelligence and compete for mates, then it needs to be accompanied by other mental facilities: a fast Theory of Mind (to converse) and social emotions (to seek high status within a group, to find a mate). So language underlies a less rational side of human nature – our irrational emotions, and the harm we may do to ourselves and others.

.

1. Introduction

The human mind is a prodigious pattern-matching engine. Throughout our lives, we learn thousands of patterns, and we rapidly retrieve them to match them to whatever we are experiencing. We think of language as part of this pattern-matching ability – matching the sounds of words, to understand what we hear. Language is seen as a benign, neutral medium for creating and expressing ideas – a wholly beneficial adaptation of the human mind. This narrow view of language is shown in figure (1a).

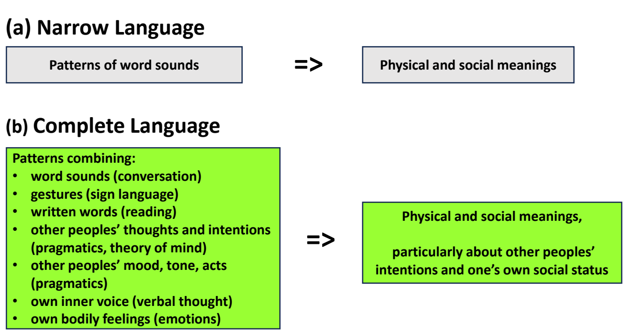


Figure 1: a narrow view of language, and a more complete view of language

This narrow view of language has arisen from the academic study of syntax and semantics – which are intellectually purer aspects of language. Looking at the many uses of language suggests a more complete view. Language is more pervasive, and language pattern matching is more wide-ranging, encompassing all the patterns of figure 1(b). Language is not just the benign, neutral medium we have taken it for; its influence is more profound, and not always beneficial. I suggest that the reasons for this lie in the evolutionary origins of language – which involve sexual selection. Understanding these origins can help us understand the role language plays in our lives, and how it is linked to a darker, less rational side of our nature[1].

1. Language Evolution and Sexual Selection

There are many theories about how human language evolved, described in previous proceedings of this conference, and in (Christiansen & Kirby, eds, 2003). These have two difficulties (Szamado & Szathmary 2006):

A. They do not account for the uniqueness of human language. If mankind has expressive language and high intelligence, why has no other species evolved a similar capability?
B. In most accounts, the fitness benefits brought by language in a natural habitat are not sufficient to offset the large metabolic costs of our expanded brains.

An account of language evolution through both sexual selection and natural selection can address these problems.

Sexual selection (Lande 1981; Maynard Smith 1982) is very widespread. It creates much of the diversity and vivid profusion of nature, such as birds' plumage or flowering plants. (Worden 2022) has proposed a hybrid account of the evolution of language, in which both natural selection and sexual selection have played a part. In this account, superior intelligence became a sexually attractive trait in *Homo Sapiens*, needed by both sexes to attract a mate (Miller 2002); and complex language evolved as the primary way to display intelligence. This hybrid account does not conflict with accounts of language evolution by natural selection. For a full account, see (Worden 2022). In short, both of the difficulties (A) and (B) are addressed by sexual selection:

.

- (A) Sexual selection leads to species-unique traits, because it acts in a unique way within each species;

[1] Due to limitations of space, some key concepts in this paper are only briefly described. By the time of the conference, a fuller version of the paper will be posted on arXiv, and on ResearchGate

- (B) Sexual selection is a process of runaway positive feedback, leading to exaggerated traits and handicaps, such as the peacock's tail, or the metabolically expensive human brain

If language evolved for the display of intelligence, to be sexually attractive and to gain high social status (in order to get a mate), some key properties of language follow:

a) It must be accompanied by high general intelligence, in order to be impressive (our enlarged brains)
b) Intelligence is displayed through conversation; the skills of conversation are a key part of language (pragmatics)
c) To impress, our speech must be fast and expressive (prodigious)
d) To impress another person in conversation, you need to know what they think, know and do not know (the Theory of Mind, or ToM)
e) You need to read their intentions though their gestures, tone of voice, and facial expressions – as well as their words.
f) To gain high status (in other peoples' eyes), requires inferring what they think about us
g) Our concept of ourselves is defined by what we think other people think about us (self-esteem)
h) To make our conversations more impressive, we rehearse them internally (verbal thought)
i) We monitor our changing self-esteem through our bodily feelings (emotions)

This helps to understand all the pattern-matching we use in language - the complete language of figure 1(b), not just the narrow language of figure 1(a). It shows that language is deeply linked to our self-esteem and emotions.

2. The Patterns of Complete Language

Research on language learning has focused on a narrow view of language - how we learn syntax and semantics, so we can understand the words we hear, and express what we mean. Syntax and semantics will not be discussed further, except to say that they use pattern-matching (technically, unification); the patterns are learnt from early childhood, and they are applied rapidly and pre-consciously. On hearing hear any utterance, we do a lot of pre-conscious pattern matching, before being consciously aware of its meaning.

I focus on the other pattern-matching in the complete language of figure 1(b). The first use of language is in conversation. To impress other people, we need to be fluent conversationalists – able to take our conversational turns within a fraction of a second (Levinson & Torreira 2015), to infer the relevance of what someone has said (Sperber & Wilson 1986), and to infer our partner's conversational intent from what they say, and from the context. These pragmatic skills require mind-reading – a Theory of Mind (ToM), to infer what the other person in a conversation may be thinking (Sperber & Wilson 2002); so that a shared cooperative intent in the conversation is part of the common ground (Stalnacker 2002, Tomasello 2014). The ToM skill may be learnt as conversational patterns, similar to the pattern learning that we use to learn the meanings of words; and the ToM is applied in conversation by fast, pre-conscious pattern matching.

So we learn a Fast Theory of Mind – an ability to infer rapidly what a conversational partner is thinking, feeling and intending, from what they say and from the context. This includes what they are thinking about our selves. The human sense of self emerges largely as a sense of 'what I think the other person is thinking about me'. This becomes the self as measured against the social norms of the group – seeing oneself in the mirror of other peoples' assessments.

The need to impress others is linked to a need to obtain high social status within a group, in order to get a mate. Our self-perceived social status is a ToM assessment of 'what I think other people think of me'. In conversation, we track that assessment, and choose what we say to maximise it. Part of this is to feel unpleasant emotions – bodily feelings that are triggered when our self-esteem is threatened - and to use those emotional feelings to guide what we say, to bolster our self-esteem when necessary. This requires us to learn the patterns of our bodily feelings arising from emotions, and to use them to guide our conversation. This fast pre-conscious pattern matching may be like the pattern matching we use to learn words and the ToM. These learnt patterns of bodily feelings are all parts of language.

In summary, complete language (figure 1a) requires us to learn thousands of complex patterns, involving word sounds, the inferred mental states of others, and our own bodily feelings. These are shown in figure 2.

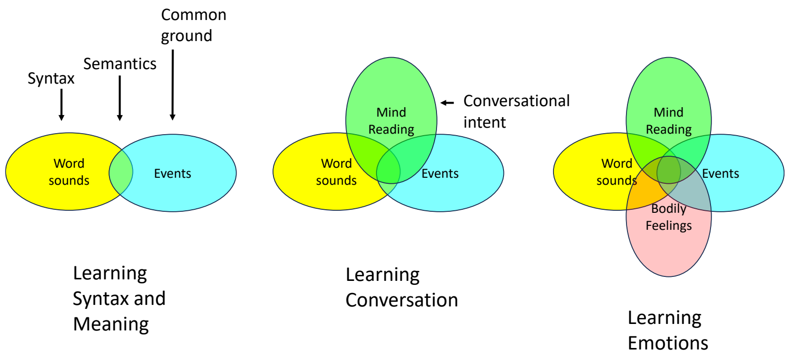


Figure 2: Three components of language learning, which are needed to use language to display intelligence.

As early as three years of age, a child learns many complex patterns, in all four quadrants of the picture (Bloom et al. 1993; Fletcher & McWhinney 1996). Human speech and thought works by fast, pre-conscious matching of these patterns. That makes the human mind a very complex dynamical system – and not always a rational one.

3. Consequences of Language for Human Nature and Rationality

To make our conversations fluent and impressive, we mentally rehearse them. This is the origin of verbal thought. As we think, we are consciously aware of the sounds of the words, and we remember them. Later recall enables us to construct extended chains of thought, and is the basis of our rationality (Pinker 2021).

When rehearsing conversations as verbal thoughts, we have in mind who the audience might be. As we think, the ToM patterns that we have learnt are matched, and we infer what the audience will think – their reactions to our words. In much verbal thought, there is a '**shadow audience**' in our minds, and we constantly infer what they will think about what we are thinking and might say. The shadow audience may be a specific person, but often it is a group, such as 'my parents' or 'the neighbours' or 'my peers at work'.

The influence of the shadow audience on our thought is pervasive:

1. The many ToM patterns which we learn in conversations match sense data (what the other person says, contextual cues). Those patterns work well enough to sustain a conversation (Levinson 1983). When the same ToM patterns are matched in our private thoughts, there is no feedback

from another person; so as a guide to what other people think, the patterns are less reliable.

2. Much of what we infer is about ourselves: ‘what my shadow audience thinks of me’; if that is negative, lower self-esteem triggers negative emotions. These are consciously felt in the body, leading to further thoughts and emotions.

So while verbal thought enables us to construct and critique long chains of reasoning, supporting our rational thought (Mercier & Sperber 2017), it also triggers self-esteem reactions though ToM patterns. These patterns, in the absence of input from others, are unreliable and irrational. Our sense of self, being based on unreliable inferences about ‘what other people will think of me’ is a second-hand and impoverished sense of ourselves – like viewing ourselves in a cracked mirror.

ToM patterns triggered by bodily feelings can lead to cascades, in which we first feel some emotion as a bodily feeling; then, using ToM patterns, we unreliably infer what our shadow audience will think of us if we show that emotion. This triggers further emotions, and further words as we try to counter negative self-esteem. These cascades may be the cause of the volatile, unpredictable, and irrational nature of human emotions. They are driven by ToM ‘shadow audience’ patterns which are learnt from an early age, and may never be un-learnt.

The need to impress other people leads to group-think and tribalism. If some opinion is held within a group, and is affirmed in conversations, then we think we will achieve high status in the group by agreeing with it. We do this in our private thoughts, which are rehearsed conversations; self-esteem is enhanced by the inferred agreement of a shadow audience. It then matters more that some opinion should agree with a group opinion, than that it fits the facts and evidence. This is group-think. Acceptance in the group is enhanced by a negative view of other groups. This leads to tribalism and rejection of out-groups, reinforced by group-think. These are some of the irrational forces that may cause people to mistreat and harm other people – man’s inhumanity to man. They start with language.

4. Conclusions

The human mind is partly rational, partly irrational. Our irrationality has caused immense harm over the ages, and continues to do so. With the growing power of technology, now more than ever we need to understand our own irrationality. This paper suggests that our rationality and irrationality both spring from the same origin – our use of language to display superior intelligence. The same fast pattern matching, which enables us to understand the words we hear, also drives our self-esteem and emotions, sometimes in harmful ways. It is a scientific priority to understand the origins of human irrationality. This paper is an attempt to do so.

References

Bloom L et al. (1993) Language development from two to three, Cambridge University Press

Christiansen M. H. and Kirby S. (Eds., 2003), *Language evolution* Oxford: Oxford University Press.

Fletcher P. and McWhinney B (1996) The handbook of child language, Blackwell, Cambridge, Mass.

Lande, R. (1981). Models of speciation by sexual selection on polygenic traits.
Proc. Natl. Acad. Sci. U. S. A. 78, 3721–3725. doi: 10.1073/pnas.78.6.3721

Levinson, S. C. (1983). Pragmatics. Cambridge, UK: Cambridge University Press.

Levinson, S. C., and Torreira, F. (2015). Timing in turn taking and its implications for processing models of language. Front. Psychol. 6:731. doi: 10.3389/fpsyg.2015.00731

Maynard-Smith, J. (1982). Evolution and the theory of games. Cambridge, UK: Cambridge University Press.

Mercier, H., and Sperber, D. (2017). The enigma of reason. Harvard University Press.

Miller, G. (2001). The mating mind: How sexual choice shaped the evolution of human nature. London, Heineman.

Pinker S. (2021) Rationality, Allen Lane, London

Sperber, D., and Wilson, D. (1986) “Relevance: communication and cognition” in Second edition (with postface) 1995 (Oxford: Blackwell)

Sperber, D., and Wilson, D. (2002) Pragmatics, modularity and mind-reading. Mind Lang. 17, 3–23. doi: 10.1111/1468-0017.00186

Stalnaker, R. (2002). Common ground. Linguist. Philosophy 25, 701–721. doi: 10.1023/A:1020867916902

Számadó, S. Z., and Szathmáry, E. (2006). Competing selective scenarios for the emergence of natural language. Trends Ecol. Evol. 21, 555–561. doi: 10.1016/j.tree.2006.06.02

Tomasello, M. (2014). A natural history of human thinking. Cambridge, Mass.: Harvard University Press.

Worden, R. (2022) The evolution of language by sexual selection, Front. Psychol., Vol 13 https://doi.org/10.3389/fpsyg.2022.1060510

Simulating Dependency Length Minimization using neural-network based learning and communication

Yuqing Zhang[1], Tessa Verhoef[2], Yuchen Lian[2], Gertjan van Noord[1], and Arianna Bisazza[1]

[*]Corresponding Author: yuqing.zhang@rug.nl
[1]Center for Language and Cognition, University of Groningen, Groningen, The Netherlands
[2]Leiden Institute of Advanced Computer Science, Leiden University, Leiden, The Netherlands

A fruitful approach to studying the influence of human cognitive biases and processes like language learning and use in shaping linguistic structure is to simulate them computationally (De Boer, 2006; Steels, 1997). Recent advances in machine learning and computational linguistics have yielded powerful (neural-network based) artificial learners that can deal surprisingly well with the complexity of human languages and can be used to set up increasingly realistic simulations (Chaabouni et al., 2021; Warstadt & Bowman, 2022). An important challenge in this line of work, however, is that such artificial learners still often behave differently from human learners (Chaabouni et al., 2019a; Galke et al., 2022).

Recently, Lian et al. (2023) proposed a novel framework for simulations of language learning and change with artificial languages and neural-network learners, which addresses some of these challenges. In the NeLLCom (Neural-agent Language Learning and Communication) framework, pairs of speaking and listening agents learn a pre-defined artificial language through supervised learning and then communicate with each other, optimizing a shared reward via reinforcement learning. Communication is simulated with a meaning reconstruction game where a speaker learns to convey a meaning m to a listener using the language it has learnt by supervised learning. This language use is tied to a shared goal: maximizing the communicative reward evaluated by the listener's prediction. Speakers are modeled as a linear-to-sequence structure whereas listeners work in the reverse direction, i.e., a sequence-to-linear structure. Sequential encoding/decoding is implemented by a recurrent neural network (RNN), specifically Gated Recurrent Units (Cho et al., 2014). In the original study, Lian et al. (2023) applied the framework to simulate the emergence of the word-order/case-marking trade-off and found that a human-like trade-off appears during communication without hard-coding specific biases in the agents.

In this work, we focus on another statistical language universal: dependency length minimization (DLM), a tendency to minimize the linear distance between heads and their dependents in natural languages (Futrell et al., 2020). Motivated by the contradictory patterns found in previous simulations (Chaabouni et al., 2019b; Zhao, 2022), we adopt NeLLCom to further investigate the minimal conditions that may lead to DLM in neural-network learners. Inspired by Fedzechkina et al. (2018)'s experiment with human learners, we expand Lian et al. (2023)'s original meaning space of agent-patient-action triplets by adding optional modifier phrases to agent and patient. Each modifier phrase consists of three items: adposition, adjective, and inanimate noun (e.g. *behind white door*). The meanings are descriptions of scenes that have only one long constituent (i.e. only subject or object has adpositional-phrase modification). Similar to NeLLCom's original setting, utterances are variable-length sequences of symbols taken from a fixed-size vocabulary: $u = [w_1, ..., w_l]$, $w_i \in V$. For each meaning, there are two possible utterance orderings (subject-object and object-subject). Ordering symbols representing short dependents closer to verbs leads to shorter total dependency length.

We train agents on a verb-initial language and a verb-final language, each comprising 50% long- and 50% short-dependency utterances. Additionally, we train on two control languages, each containing only short or long dependency utterances, respectively. We then conduct evaluations on meanings unseen during training. In this initial setup, speakers do not regularize towards reducing DL in production nor do the shorter-dependency languages show a learning advantage compared to their longer-dependency counterparts.

We then consider three additional factors to make the simulation more realistic: (i) introducing noise during listening (Futrell & Levy, 2017) through a word dropout technique (Gal & Ghahramani, 2016), (ii) modeling non-uniform word distributions and selectional preferences of verbs (McRae et al., 1998), i.e. strength of association of one action with one agent or patient, and (iii) testing listeners' incremental utterance processing (Kamide et al., 2003), or the extent to which an utterance's meaning can be guessed before hearing it entirely. We find that, during communication, neural learners tend to regularize towards one word order instead of one dependency length, failing to display a DLM preference in their productions. However, the proposed factors contribute to a small but consistent learning advantage of shorter dependencies for listening agents of the verb-initial language, but not for the verb-final language, which is consistent with patterns found in natural languages (Jing et al., 2022). Specifically, for the verb-initial language, we find that: 1) under noisy conditions, listeners learn the short-dependency language slightly better and faster than the long-dependency one; 2) the presence of noise affects the learning accuracy of uniform languages more severely than languages with skewed conditional word distributions; 3) the short-dependency language shows an advantage over the long one when evaluated incrementally, suggesting that the disambiguation of meanings related to local dependencies tends to occur earlier in the sentence.

Future directions include integrating incremental processing modeling into our use of the NeLLCom framework for investigating language emergence.

References

Chaabouni, R., Kharitonov, E., Dupoux, E., & Baroni, M. (2019a). Anti-efficient encoding in emergent communication. *Advances in Neural Information Processing Systems*, *32*, 6293–6303.

Chaabouni, R., Kharitonov, E., Dupoux, E., & Baroni, M. (2021). Communicating artificial neural networks develop efficient color-naming systems. *Proceedings of the National Academy of Sciences*, *118*(12), e2016569118.

Chaabouni, R., Kharitonov, E., Lazaric, A., Dupoux, E., & Baroni, M. (2019b). Word-order biases in deep-agent emergent communication. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 5166–5175). Florence, Italy: Association for Computational Linguistics.

Cho, K., Merriënboer, B. van, Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1724–1734). Doha, Qatar: Association for Computational Linguistics.

De Boer, B. (2006). Computer modelling as a tool for understanding language evolution. In *Evolutionary epistemology, language and culture: A non-adaptationist, systems theoretical approach* (pp. 381–406). Springer.

Fedzechkina, M., Chu, B., & Florian Jaeger, T. (2018). Human information processing shapes language change. *Psychological science*, *29*(1), 72–82.

Futrell, R., & Levy, R. (2017). Noisy-context surprisal as a human sentence processing cost model. In *Proceedings of the 15th conference of the european chapter of the association for computational linguistics: Volume 1, long papers* (pp. 688–698). Valencia, Spain.

Futrell, R., Levy, R. P., & Gibson, E. (2020). Dependency locality as an explanatory principle for word order. *Language*, *96*(2), 371–412.

Gal, Y., & Ghahramani, Z. (2016). A theoretically grounded application of dropout in recurrent neural networks. *Advances in neural information processing systems*, *29*.

Galke, L., Ram, Y., & Raviv, L. (2022). Emergent communication for understanding human language evolution: What's missing? In *Emergent communication workshop at iclr 2022.* virtual.

Jing, Y., Blasi, D. E., & Bickel, B. (2022). Dependency-length minimization and its limits: A possible role for a probabilistic version of the final-over-final condition. *Language*, *98*(3).

Kamide, Y., Altmann, G. T., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and language*, *49*(1), 133–156.

Lian, Y., Bisazza, A., & Verhoef, T. (2023). Communication Drives the Emergence of Language Universals in Neural Agents: Evidence from the Word-order/Case-marking Trade-off. *Transactions of the Association for Computational Linguistics*, *11*, 1033-1047.

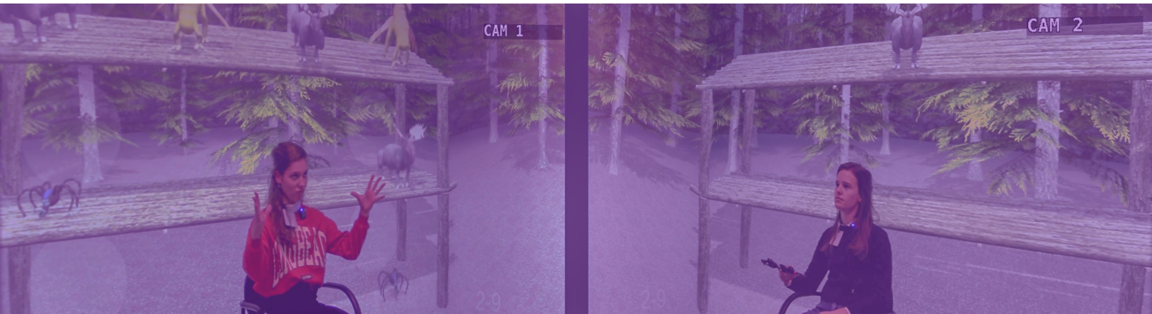
McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, *38*(3), 283–312.

Steels, L. (1997). The synthetic modeling of language origins. *Evolution of communication*, *1*(1), 1–34.

Warstadt, A., & Bowman, S. R. (2022). What artificial neural networks can tell us about human language acquisition. *Algebraic Structures in Natural Language*, 17–60.

Zhao, Y. (2022). *How to probe linguistic knowledge and bias.* Unpublished doctoral dissertation, The University of Arizona.

Author Index



THE EVOLUTION of LANGUAGE

In these proceedings, we collected 129 papers and abstracts from the 15th International Conference on the Evolution of Language (Evolang XV) held in Madison, Wisconsin, USA, in May 2024.

As the leading international event on the evolution of language, Evolang promotes a multidisciplinary approach to the study of language and communication systems. The contributions of this volume are written by scholars from all continents and from various fields which include biology, psychology, archaeology, genetics, AI, neuroscience, philosophy and linguistics, among many others.