

# Induction and interaction in the evolution of language and conceptual structure

Jon W. Carr

2019

Submitted in fulfilment of the degree of Doctor of Philosophy to  
*School of Philosophy, Psychology and Language Sciences*  
*University of Edinburgh*



When the lord, also known as god, realised that adam and eve, although perfect in every outward aspect, could not utter a word or make even the most primitive of sounds, he must have felt annoyed with himself, for there was no one else in the garden of eden whom he could blame for this grave oversight, after all, the other animals, who were, like the two humans, the product of his divine command, already had a voice of their own, be it a bellow, a roar, a croak, a chirp, a whistle or a cackle. In an access of rage, surprising in someone who could have solved any problem simply by issuing another quick fiat, he rushed over to adam and eve and unceremoniously, no half-measures, stuck his tongue down the throats of first one and then the other. From the texts which, over the centuries, have provided a somewhat random record of those remote times, be it of events that might, at some future date, be awarded canonical status and others deemed to be the fruit of apocryphal and irredeemably heretical imaginations, it is not at all clear what kind of tongue was being referred to here, whether the moist, flexible muscle that moves around in the buccal cavity and occasionally outside it too, or the gift of speech, also known as language, that the lord had so regrettably forgotten to give them and about which we know nothing, since not a trace of it remains, not even a heart engraved in the bark of a tree, accompanied by some sentimental message, something along the lines of I love eve.

— José Saramago (2011)





# Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified. The papers incorporated into this thesis are my own work and are included with permission from the copyright holders and with due acknowledgement made to my coauthors:

**Paper 1** Carr, J. W., & Smith, K. (2016). Modeling language transmission. In T. K. Shackelford & V. A. Weekes-Shackelford (Eds.), *Encyclopedia of evolutionary psychological science*. Springer. doi:10.1007/978-3-319-16999-6\_3353-1

**Paper 2** Carr, J. W., Smith, K., Culbertson, J., & Kirby, S. (2018, July 1). Simplicity and informativeness in semantic category systems. Preprint submitted to *PsyArXiv*. doi:10.31234/osf.io/jkfyx

**Paper 3** 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, 892–923. doi:10.1111/cogs.12371

A handwritten signature in black ink, appearing to read 'Jon Carr', with a stylized, cursive script.

Jon Carr

29 August 2018



# Summary

This thesis explores how learning a language and using it to communicate affect the structure and organization of concepts, such as colour, kinship, and spatial relationship categories. The thesis begins by outlining some theoretical perspectives on learning and communication in relation to language and concepts. These perspectives are then formalized in a computational model of an idealized language learner, and the predictions of this model are tested experimentally. The results indicate that the process of learning is best understood in terms of a cognitive preference for simplicity. Moreover, when a language is passed on from one generation of learner to the next, this simplicity preference is amplified, causing the language to degenerate in its ability to express useful conceptual distinctions. However, if the language is also used to communicate with others, this process of degeneration is halted, and the language instead develops a special type of structure – compositionality – which allows it to be both simple and communicatively useful at the same time. The thesis therefore provides insight into the principal forces that have shaped the structural properties of language.



# Abstract

Languages evolve in response to various pressures, and this thesis adopts the view that two pressures are especially important. Firstly, the process of learning a language functions as a pressure for greater simplicity due to a domain-general cognitive preference for simple structure. Secondly, the process of using a language in communicative scenarios functions as a pressure for greater informativeness because ultimately languages are only useful to the extent that they allow their users to express – or indeed represent – nuanced meaning distinctions. These two fundamental properties of language – simplicity and informativeness – are often, but not always, in conflict with each other. In general, a simple language cannot be informative and an informative language cannot be simple, resulting in the simplicity–informativeness tradeoff. Typological studies in several domains, including colour, kinship, and spatial relations, have demonstrated that languages find optimal solutions to this tradeoff – optimal solutions to the problem of balancing, on the one hand, the need for simplicity, and on the other, the need for informativeness.

More specifically, the thesis explores how inductive reasoning and communicative interaction contribute to simple and informative structure respectively, with a particular emphasis on how a continuous space of meanings, such as the colour spectrum, may be divided into discrete labelled categories. The thesis first describes information-theoretic perspectives on learning and communication and highlights the fact that one of the hallmark features of conceptual structure – which I term compactness – is not subject to the simplicity–informativeness tradeoff, since it confers advantages on both learning and use. This means it is unclear whether compact structure derives from a learning pressure or from a communicative pressure. To complicate matters further, some researchers view learning as a pressure for simplicity, as outlined above, while

others have argued that learning might function as a pressure for informativeness in the sense that learners might have an *a-priori* expectation that languages ought to be informative.

The thesis attempts to resolve this by formalizing these different perspectives in a model of an idealized Bayesian learner, and this model is used to make specific predictions about how these perspectives will play out during individual concept induction and also during the evolution of conceptual structure over time. Experimental testing of these predictions reveals overwhelming support for the simplicity account: Learners have a preference for simplicity, and over generational time, this preference becomes amplified, ultimately resulting in maximally simple, but nevertheless compact, conceptual structure. This emergent compact structure remains limited, however, because it only permits the expression of a small number of meaning distinctions – the emergent systems become degenerate.

This issue is addressed in the second part of the thesis, which compares the outcomes of three experiments. The first replicates the finding above – compact categorical structure emerges from learning; the second and third experiments compare artificial and genuine pressures for expressivity, and they show that it is only in the presence of a live communicative task that higher level structure – a kind of statistical compositionality – can emerge. Working together, the low-level compact categorical structure, derived from learning, and the high-level compositional structure, derived from communicative interaction, provide a solution to the simplicity–informativeness tradeoff, expanding on and lending support to various claims in the literature.

# Acknowledgements

This thesis wouldn't exist were it not for the guidance, patience, and compassion of three people: Simon Kirby, Kenny Smith, and Jenny Culbertson. Simon has the miraculous ability to find structure in my chaotic thoughts – he is a very efficient compressor – and he often understands the full ramifications of my work months or even years before I do, knowing exactly the right moment to plant exactly the right seed in my mind. Simon, I am so happy you are here to see this through with me. Kenny is unmatched in his technical knowledge, attention to detail, and ability to rapidly grasp complex issues; the work presented herein has benefited substantially from his detailed and perceptive feedback. And Jenny is a wonderfully grounding presence; passionate and thoughtful in equal measure, she has taught me two important sentences: *What's the hypothesis!?* and *Help me understand this!*

I will be lost without them.

Andrew Perfors and Chris Cummins were wonderful examiners; insightful, friendly, and genuinely curious, they proved that a defence really is what everyone told me it would be: a nice chat about my work.

I am indebted to so many other people. First and foremost, to members of the *Centre for Language Evolution* (née *Language Evolution and Computation Research Unit*) who I've had the pleasure to work alongside, every single one of whom has, at one point or another, given me valuable advice or changed the way I think about some issue: Mark Atkinson, Richard Blythe, Fausto Carcassi, Hannah Cornish, Chrissy Cuskley, Isabelle Dautriche, Olga Fehér, Vanessa Ferdinand, Molly Flaherty, Stella Frank, Jim Hurford, Tamar Johnson, Hanna Järvinen, Jasmeen Kanwal, Andres Karjus, Fiona Kirton, Jia Loy, Alex Martin, Yasamin Motamedi, Alan Nielson, Jonas Nölle, Cathleen O'Grady,

Nina Poth, Hugh Rabagliati, Carmen Saldaña, Asha Sato, Marieke Schouwstra, José Segovia Martin, Cat Silvey, Matt Spike, Kevin Stadler, Mónica Tamariz, Bill Thompson, Rob Truswell, Svenja Wagner, James Winters, and Marieke Woensdregt.

Secondly, to my office mates who have contributed to a delightful working environment: Soundess Azzabou-Kacem, Michela Bonfieni, Adam Clark, Steph DeMarco, E Jamieson, Daniel Lawrence, Julie-Anne Meaney, Katerina Pantoula, Steve Rapaport, Jade Sanstead, George Starling, Ellise Suffill, and Tom Wood.

Thirdly, to the members of faculty who have influenced my thinking in one way or another – Holly Branigan, Josef Fruehwald, Heinz Giegerich, Nikolas Gisborne, Pavel Iosad, James Kirby, Martin Pickering, Korin Richmond, and Hannah Rohde – and members of the support staff who have ensured that my time in Edinburgh has been so exceedingly smooth: Alejandro Alonso Gonzalo, Julie Anderson, Stephen Boyd, Lynsey Buchanan, Stephanie Fong, Susan Hermiston, Katie Keltie, Alistair Kirkhope, Cedric Macmartin, Steven McGauley, Michael Murray, Toni Noble, Alisdair Tullo, Becky Verdon, Davie Wilkinson, and – no doubt – very many others behind the scenes.

And finally, to the remarkable people I've had the pleasure to meet in various corners of the globe: Alex Carstensen, Bart de Boer, Morton Christiansen, Nico Claidière, Davide Crepaldi, Dan Dediu, Michael Dunn, Joël Fagot, Nick Fay, Peter Gärdenfors, Robert Hawkins, Niklas Johansson, Anna Jon-And, Vera Kempe, Hannah Little, Luke McCrohon, Ashley Miklos, Thomas Müller, Kaz Okanoya, Andrea Ravignani, Limor Raviv, Terry Regier, Gareth Roberts, Seán Roberts, Justin Sulik, Tessa Verhoef, Andy Wedel, and – who could forget? – Bodo Winter.

To you all, I am deeply grateful.

I also want to thank the *Scottish Graduate School of Social Science*, who funded this research through a block grant from the *Economic and Social Research Council* (grant number ES/J500136/1). Two non-human entities, and their human overlords, also deserve special thanks: eddie and blake, CTRL-D for now.

And finally, I am so immensely thankful to Wil, who has always believed in me and put up with a constant lack of attention for several years now, often keeping me alive while my eyes were glued to this goddamned screen for days on end. For what it's worth, this is dedicated to you.



# Contents

1	Introduction	1
1.1	Preface to Paper 1	2
	Paper 1: <i>Modeling language transmission</i>	3–7
	Introduction	3
	Cultural evolution and language	3
	Cultural transmission gives rise to compositionality	4
	Compositionality depends on transmission <i>and</i> communication	4
	Cultural transmission gives rise to semantic categories	6
	Conclusion	6
1.2	Summary of Paper 1	8
1.3	On Iterated Learning	9
1.4	Roadmap	11
2	Categorization, Compression, and Communication	15
2.1	Concepts and Categorization	16
2.1.1	Words, concepts, kinds, and categories	16
2.1.2	Convexity and compactness	17
2.1.3	Iterated learning and conceptual structure	20
2.2	Simplicity and Learning	21
2.2.1	Bayesian inference as a model of learning	22
2.2.2	Coding and probability: Some preliminaries	23
2.2.3	Sources of compression in language structure	25
2.2.4	Algorithmic probability and the minimum description length principle	27

2.2.5	Complexity and concept learning	30
2.2.6	Hallmarks of simple category systems	32
2.3	Informativeness and Communication	35
2.3.1	Informativeness in typological datasets	36
2.3.2	Communicative cost	38
2.3.3	Hallmarks of informative category systems	43
2.4	The Simplicity–Informativeness Tradeoff	45
2.5	Conclusion to Chapter 2	47
3	Simplicity from Induction	49
3.1	Preface to Paper 2	50
	Paper 2: <i>Simplicity and informativeness in semantic category systems</i>	52–86
	Introduction	52
	Model	56
	Method	56
	Results	64
	Summary	66
	Experiment 1: Category learning	67
	Method	68
	Results	72
	Summary	72
	Experiment 2: Iterated learning	73
	Method	73
	Results	74
	Model fit	76
	Summary	80
	Discussion	80
	Conclusion	82
3.2	Summary of Paper 2	87
3.3	Simulating Participant Communication	90
3.4	Complexity and the Rectangle Code	91
3.4.1	Deviations from Fass & Feldman’s rectangle code	92
3.4.2	Rectangular decomposition	93

3.5	Metropolis–Hastings and the Proposal Function	97
3.5.1	Cell mutation	98
3.5.2	Rectangle mutation	101
3.6	Testing the Model Fit Procedure	101
3.7	Conclusion to Chapter 3	104
4	Informativeness from Interaction	107
4.1	Preface to Paper 3	108
	Paper 3: <i>The cultural evolution of structured languages in an open-ended, continuous world</i>	110–141
	Introduction	110
	Experiment 1: Basic transmission	114
	Method	114
	Results	117
	Summary	122
	Experiment 2: Transmission with an artificial expressivity pressure	123
	Method	123
	Results	124
	Summary	125
	Experiment 3: Transmission with communication	126
	Method	127
	Results	128
	Summary	131
	Discussion	132
	Conclusion	135
	Appendix A: Online dissimilarity rating task	139
	Appendix B: Dissimilarity judgments between target and selected triangles in Experiment 3	141
4.2	Summary of Paper 3	142
4.3	Structure and the Mantel Test	143
4.4	Problems with Page’s Test: Some Errata	145
4.5	Simplicity and Informativeness	148
4.6	Conclusion to Chapter 4	149

---

5	Conclusion	153
5.1	Recapitulation	154
5.2	Future directions	156
5.3	Final thoughts	157
	Appendices	159
A	Paper 2, Supplement S2: All model results	159
B	Paper 2, Supplement S3: Participant exclusion and attrition	177
C	Paper 2, Supplement S4: Individual participant results in Experiment 1	181
D	Paper 3, Supplement S1: Experimental briefs	189
E	Paper 3, Supplement S2: Geometric measure of triangle dissimilarity	193
F	Paper 3, Supplement S3: MDS plots for all generations in all chains	199
	References	213





# Chapter 1

## Introduction

A struggle for life is constantly going on amongst the words and grammatical forms in each language. The better, the shorter, the easier forms are constantly gaining the upper hand, and they owe their success to their own inherent virtue.

— Charles Darwin (1871)<sup>1</sup>

The idea that languages evolve by Darwinian means is an apt starting place for this thesis, which is firmly rooted in an evolutionary approach to the language sciences. And the words ‘better’, ‘shorter’, and ‘easier’ are especially fitting given its main argument – that languages evolve to become *better* adapted to the communicative needs of their users, while also evolving *shorter* grammars, making them *easier* to learn. Paraphrasing Dobzhansky (1973), it has been remarked that ‘nothing in linguistics makes sense except in the light of evolution’ (Kirby, 2014). However, perhaps more so than biology, the emergence and evolution of language is inherently difficult to study due to its ephemeral nature as an ever-changing, non-ossifying cultural behaviour.

The most recent wave of scientific research into the evolution of language began a few decades ago with pioneering work using agent-based computational simulations (e.g. Brighton, 2002; Hurford, 1989; Kirby, 2002; Nowak & Krakauer, 1999; Oliphant, 1996; Smith, 2004; Steels, 1995). More recently, experimental models have also become

---

<sup>1</sup> In fact, these words are Darwin’s (1871) paraphrase of a few sentences in Müller (1870), who was reviewing an English translation by Bikkers (1869) of a German text by Schleicher (1863). It is, of course, highly satisfying that a quotation about language evolution has itself undergone several generations of selection and recombination. See Dingemanse (2013) for a brief history of this quote.

prevalent (e.g. Galantucci, 2005; Garrod, Fay, Lee, Oberlander, & MacLeod, 2007; Little, Eryilmaz, & de Boer, 2017; Perlman, Dale, & Lupyan, 2015; Scott-Phillips, Kirby, & Ritchie, 2009; Selten & Warglien, 2007), early precursors to which can be found in the works of, for example, Bartlett (1932) and Esper (1966).

I believe that the combination of models and experiments, underpinned by strong theoretical claims, offers the best way to make progress in the field, and herein I attempt to do precisely that. A central premise of this thesis is that simple computational principles can, at least in part, explain the high-level complexity we observe in human behaviours and therefore the cultural phenomena that arise from such behaviours. As such, I believe there are three main entry points to discovering what it means to be human: computational principles (with a special emphasis on learning), human behaviour (with a special emphasis on categorization), and cultural phenomena (with a special emphasis on language).

The thesis consists of five chapters, three of which (Chapters 1, 3, and 4), are built around a paper. In this introductory chapter, I briefly review some of the background on which this thesis rests and provide some high-level definitions of key terms. This material will then be expanded upon in Chapter 2, which provides a much more rigorous theoretical background to support the primary content presented in Chapters 3 and 4.

## 1.1 Preface to Paper 1

Paper 1 has been accepted for publication in the *Encyclopedia of Evolutionary Psychological Science*, which is scheduled for release in 2019. The advance online version (Carr & Smith, 2016) is reproduced over the subsequent pages with permission from the publisher. I wrote the paper, created the figures, and handled the submission process. Kenny Smith gave advice on the structure of the paper and edited it. The citations may be looked up on page 7 or in the references list at the end of this volume. In particular, note that ‘Carr et al. (2016)’ refers to Paper 3 from this thesis, which I will return to in Chapter 4.



---

# M

---

## Modeling Language Transmission

Jon W. Carr and Kenny Smith  
School of Philosophy, Psychology and Language  
Sciences, University of Edinburgh, Edinburgh,  
UK

### Definition

Languages adapt as they are transmitted from one generation to the next. Modeling language transmission in computer simulations and laboratory experiments shows how this process gives rise to the structure found in language.

### Introduction

Language is a defining characteristic of our species, so understanding its evolutionary origins is central to understanding human evolution. In their seminal paper, Pinker and Bloom (1990) argued that the evolution of language is best understood as the result of conventional Darwinian processes, just like other complex biological traits. However, languages themselves also adapt and evolve over repeated episodes of learning and use, providing two evolutionary mechanisms that shape language: the biological evolution of the human capacity for language *and* the cultural evolution of language itself. This entry outlines the consequences of cultural evolution for language and

gives examples of how modeling language transmission can shed light on how language evolved.

## Cultural Evolution and Language

Like many other human behaviors, language is socially learned and culturally transmitted: Humans learn the language of their speech community by observing the linguistic behaviors of other members of that community. More specifically, languages are transmitted via *iterated learning*: A language is learned by observing the linguistic behavior of another individual who learned their language in the same way. That humans are able to learn language presumably reflects some cognitive capacity or combination of capacities that is unique to humans (Hauser et al. 2002). However, repeated episodes of learning and use also allow for the cultural evolution of languages: Linguistic variants that are difficult to learn, impose substantial processing burdens, or do not meet the communicative needs of language users will tend to be replaced by those that are more learnable, easier to process, or more functional (Christiansen and Chater 2008). This is because the mistakes that language users make during learning, and the modifications they make while communicating, tend to be in favor of more learnable, more functional forms; poorly adapted variants will be replaced by superior ones. Cultural evolution thus gives rise to languages that are

well adapted to being transmitted from one generation to the next.

Mathematical, computational, and experimental techniques developed over the past two decades have made it possible to systematically investigate how cultural processes shape language (see Kirby et al. 2014 for a review). This line of research has demonstrated that some of the fundamental properties of language can be explained as products of cultural evolution, thus highlighting the importance of understanding the role of culture in explaining language design and reframing the debate on the biological evolution of the language faculty (Thompson et al. 2016). This entry reviews some of this work here, focusing on experimental models of the emergence of compositional and categorical structure in language.

### Cultural Transmission Gives Rise to Compositionality

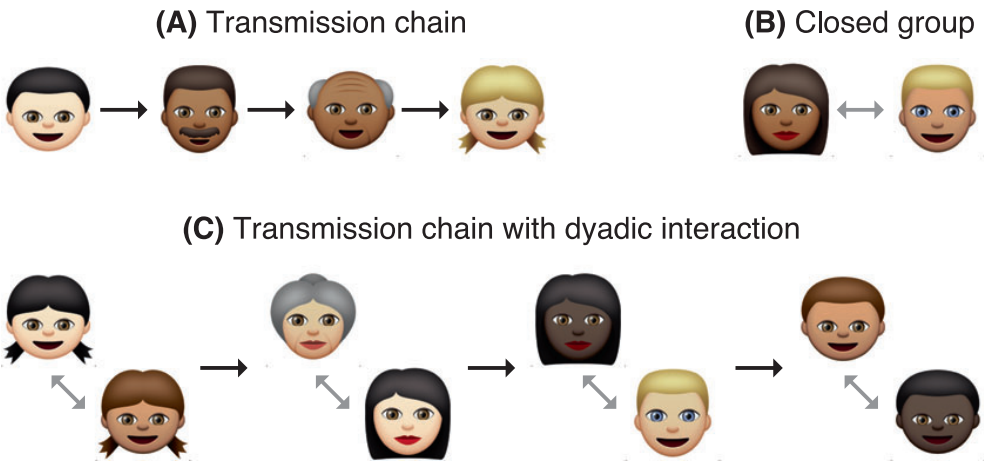
Language is compositional: The meaning of a complex utterance is a function of the meaning of its parts and the order in which those parts are combined. By combining a set of linguistic units in a particular order (e.g., *the dog bit the man*), a language user is able to form a complex meaning that is systematically related to an utterance that uses a different set of words (e.g., *the cat bit the man*) or that places the words in a different order (e.g., *the man bit the dog*). This compositional structure is central to the expressive power of language; with knowledge of the linguistic units and rules of combination, language users are able to produce and understand any complex utterance—even those that have never been encountered before.

In the first work of its kind, Kirby et al. (2008) ran an experiment showing that the property of compositionality can emerge as a result of language transmission, replicating the results of earlier computer models (e.g., Kirby 2002). Participants had to learn an “alien” language which consisted of words for colored moving shapes. After a training phase in which participants observed the objects together with their

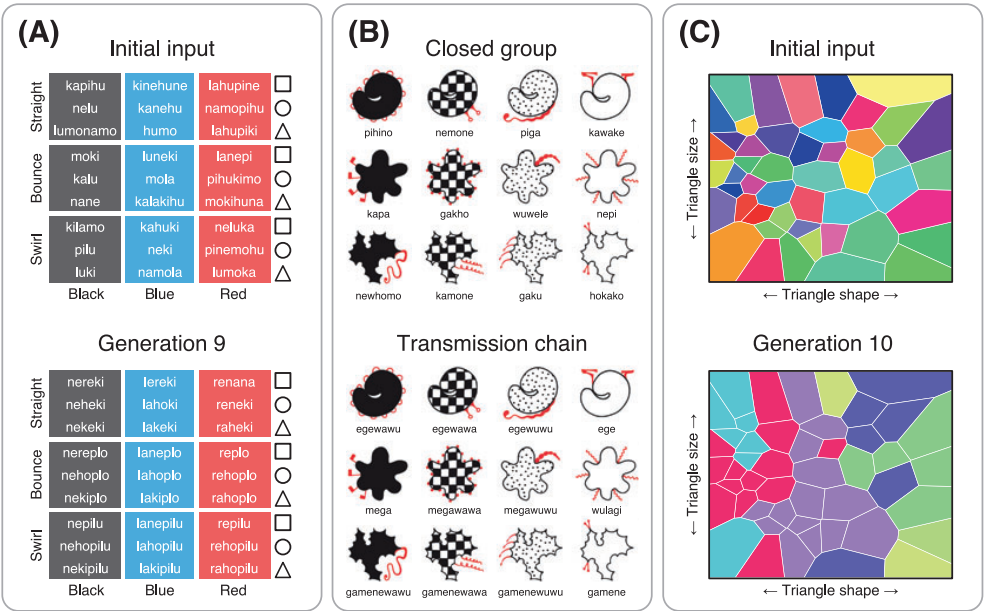
labels, participants were prompted to recall the labels for those objects. The responses of a given participant were then taught to a new participant, whose responses were in turn taught to another new participant, thus modeling what happens when languages are transmitted between individuals (Fig. 1a). Each transmission chain was initialized with an unstructured, non-compositional language in which every object was associated with a randomly generated label. After around ten generations of the iterated learning process, linguistic systems emerged that exhibited compositional structure. An example of this result is shown in Fig. 2a. The initial input language taught to the first participant in a chain contains no system-wide structure, but by the ninth generation, the language had evolved a compositional system in which the first syllable encodes color, the second syllable encodes shape, and the final syllable encodes movement.

### Compositionality Depends on Transmission and Communication

Languages are not merely transmitted from person to person via learning and recall; they are used for communication, and the communicative use of language provides the input to language learning. This means that language is shaped by two pressures. On the one hand, a language needs to be expressive—it should allow its users to convey important distinctions when communicating. On the other hand, it also needs to be learnable. These pressures for expressivity and learnability are not necessarily aligned: Languages that convey many distinctions are likely to be harder to learn than languages that encode few distinctions. Indeed, the easiest language to learn would be one in which every concept was conveyed by a single, maximally ambiguous utterance, but such a language would be inexpressive. Kirby et al. (2008), described above, used an artificial proxy for expressivity: If a participant provided the same label for two objects, only one of those labels was passed on to the next learner, thus concealing evidence that languages could be inexpressive. In another experiment in the same paper, Kirby



**Modeling Language Transmission, Fig. 1** Three models of cultural transmission. (a) shows a simple transmission chain in which a language is passed from one individual to another. (b) shows a pair of language users who interact back and forth using a language. (c) shows a transmission chain with dyadic interaction at each generation



**Modeling Language Transmission, Fig. 2** Results from three iterated learning experiments. (a) shows results from experiment 2 from Kirby et al. (2008). The initial input language lacks systematic structure, but after nine generations of cultural transmission, the language evolves a compositional system. (b) shows results from Kirby et al. (2015). Under the closed-group method, a holistic language emerges; under the chain method, a compositional language emerges. (c) shows results from Carr et al. (2016). The initial input language contains no categories (each color is a different word), but after ten generations of cultural transmission, the continuous meaning space is carved up into semantic categories

et al. (2008) found that removing this artificial pressure for expressive languages produced a radically different outcome: Rather than becoming compositional, the languages are rapidly simplified, losing words and distinctions at every episode of transmission.

In a follow-up series of computer models and experiments, Kirby et al. (2015) explored the trade-off between these competing pressures, modeling the expressivity pressure in a more naturalistic way by having participants use the language they had learned in a communication game. In one condition, two speakers had to communicate back and forth about a small set of objects (Fig. 1b). After interacting for some time, the pair of language users developed a system in which each object was described by a unique, idiosyncratic word—communicatively functional but lacking compositional structure (as shown in Fig. 2b). This condition was referred to as a *closed group*, since unlike Kirby et al. (2008) no new, naïve participants were introduced. In comparison, in the *transmission chain* condition, two participants had to communicate about the same objects, but the language was then passed on from one pair of participants to another: The language produced during communication by one pair became the input to learning for the next pair (Fig. 1c). This combination of cultural transmission to naïve learners (imposing a pressure for learnability) plus communication (favoring expressivity) led to languages with compositional structure (as shown in Fig. 2b). Communication alone, or learning alone, is not sufficient to drive the evolution of compositional structure; instead, compositionality is language's solution to pressures requiring it to be as simple and as learnable as possible without sacrificing expressive power.

### Cultural Transmission Gives Rise to Semantic Categories

As described above, languages combine linguistic units, such as words, according to a compositional system. These units pick out *categories* rather than individual items or actions. For example, in English, the space of possible drinking vessels is

carved up by a small number of words (e.g., *bottle*, *cup*, *flask*, *glass*, and *mug*). This categorical structure allows language users to refer to an infinite range of possible meanings using a manageable, finite number of labeled categories; this, in combination with compositional structure, is fundamental to the communicative power of human languages.

Carr et al. (2016) show how this categorical structure develops through iterated learning. Previous experiments (including Kirby et al. 2008, 2015) had participants learn and communicate about meanings drawn from a small, finite set. Carr et al. (2016) instead introduced a continuous and open-ended meaning space. Participants had to learn words for, and subsequently label, triangles that were randomly generated by selecting three vertices on a plane, such that there were effectively infinitely many objects participants could be faced with. In addition, participants were always tested on their ability to label entirely novel triangles, none of which they had seen during training. After ten generations of iterated learning using the transmission chain paradigm (Fig. 1a), category systems emerged in which this continuous space of possible triangles was carved up into around four or five categories that related primarily to their shape and size (as shown in Fig. 2c). When this experiment was adapted to include a communication game at each generation (Fig. 1c), as in Kirby et al. (2015), this combination of pressures for expressivity and learnability resulted in emergent languages that exhibited both categorical and compositional structure, thus demonstrating that both semantic categories and compositional structure can arise simultaneously out of cultural evolutionary processes.

### Conclusion

Humans are the only known species with a communication system as complex as language, which must reflect unique features of our biological endowment (and thus our unique evolutionary history). Biology can provide an explanation for the basic building blocks required for language, such as the capacity for vocal learning found in

other animals or the capacity and motivation to reason about the mental states and communicative intentions of others (Fitch 2010). Nevertheless, it is clear that cultural processes play a potentially important role in explaining the structure of human language. These processes can be studied in the lab, and a growing number of experiments that model what happens to languages when they are transmitted across generations have shown that at least some of the universal properties of language can be explained as a product of cultural evolution. This suggests that biological evolution should be seen as providing the basis on which cultural evolution can operate, with the detailed structural properties of language being a product of cultural evolution. Modeling language transmission therefore has an important role to play in helping us understand how language evolved.

### Cross-References

- Communication
- Communication and Social Cognition
- Darwin on the Origin of Language
- Evolution of Culture
- Language
- Language Acquisition
- Language Development
- Language Instinct, The
- Language Modularity
- Language Preadaptations
- Learning
- Laryngeal Descent
- Linguistic Evolution
- Meaning (Philosophy)
- Mother Tongue Hypothesis
- Musical Protolanguage
- Non-Human Vocal Communication
- Pinker's (1994) the Language Instinct
- Social Learning and Social Cognition

- Symbolic Culture
- Transmitted Culture
- Universal Grammar

### References

- Carr, J. W., Smith, K., Cornish, H., & Kirby, S. (2016). The cultural evolution of structured languages in an open-ended, continuous world. *Cognitive Science*. doi:10.1111/cogs.12371.
- Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, 31, 489–558. doi:10.1017/S0140525X08004998.
- Fitch, W. T. (2010). *The evolution of language*. Cambridge, UK: Cambridge University Press.
- 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, 1569–1579. doi:10.1126/science.298.5598.1569.
- Kirby, S. (2002). Learning, bottlenecks and the evolution of recursive syntax. In T. Briscoe (Ed.), *Linguistic evolution through language acquisition: Formal and computational models* (pp. 173–203). Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511486524.006.
- 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 of the United States of America*, 105, 10681–10686. doi:10.1073/pnas.0707835105.
- Kirby, S., Griffiths, T. L., & Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, 28, 108–114. doi:10.1016/j.conb.2014.07.014.
- Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141, 87–102. doi:10.1016/j.cognition.2015.03.016.
- Pinker, S., & Bloom, P. (1990). Natural language and natural selection. *Behavioral and Brain Sciences*, 13, 707–784. doi:10.1017/S0140525X00081061.
- Thompson, B., Kirby, S., & Smith, K. (2016). Culture shapes the evolution of cognition. *Proceedings of the National Academy of Sciences of the United States of America*, 113, 4530–4535. doi:10.1073/pnas.1523631113.

## 1.2 Summary of Paper 1

Paper 1 contains four key messages that are foundational to this thesis, and it is worth restating them here. The first is that human biology provides the underpinnings for language – language could not happen without whatever unique contribution our biology provides – but, equally, languages themselves are socially learned, used, and transmitted, and their structure develops in response to these processes. In particular, this thesis places special emphasis on the process of *iterated learning*, which we defined as:

**Iterated learning:** A language is learned by observing the linguistic behaviour of another individual who learned their language in the same way.

I will have more to say about this momentarily.

The second key message was a statement of the position taken by Kirby, Tamariz, Cornish, and Smith (2015), who argue that language structure emerges when two opposing forces are in play: The pressure from learning, which acts to make languages more ‘compressible’, and the pressure from communication, which acts to make languages more ‘expressive’. This is also the view generally adopted in this thesis, although I tend to prefer slightly different terminology, often using the terms *simple* rather than *compressible* and *informative* rather than *expressive*. This change of terminology is primarily motivated by a desire to unify the view from Kirby and colleagues with another view from Regier and colleagues, which we will come to in the next chapter.

The third key message was that the structural properties of language emerge as solutions to these two competing pressures. In particular, Kirby, Cornish, and Smith (2008) demonstrated that compositional structure – a hallmark property of human language – can emerge through iterated learning, but only when an ‘artificial expressivity pressure’ is included. Kirby et al. (2015) later demonstrated the same finding using a communicative task as the expressivity pressure. The reason compositional structure emerges under these conditions is because it is both simple/compressible (only a comparatively small number of linguistic units need to be learned) but also informative/expressive (the units may be combined to express many different meanings). Compositionality thus satisfies the two pressures.

The fourth key message of the paper was that the emergence of compositional structure is dependent on preexisting semantic categories; the participants must agree on

what these categories are in order to label them according to a compositional system. In Kirby et al. (2008, 2015), these semantic categories (colours, shapes, etc.) are already established – participants are already aware of a distinction between say RED and BLUE, so the task becomes one of mapping compositional units onto these established semantic categories. This is where this thesis comes in. Chapter 3 is concerned with how such semantic categories are established in the first place or, more specifically, how (iterated) learning shapes the structure of semantic category systems. Then, Chapter 4 explores how both semantic categories and compositional structure can emerge together.

### 1.3 On Iterated Learning

Aside from the emergence of compositionality highlighted by Kirby et al. (2008, 2015), iterated learning has been studied in a wide variety of domains that relate not just to language but also to other culturally transmitted behaviours. These domains have included: combinatorial structure (del Giudice, 2012; Verhoef, Kirby, & de Boer, 2015), context (Tinits, Nölle, & Hartmann, 2017; Winters, Kirby, & Smith, 2015), the manual modality (Motamedi, Schouwstra, Smith, Culbertson, & Kirby, under review), music (Ravignani, Delgado, & Kirby, 2016), sound symbolism (Johansson, Carr, & Kirby, in prep.), regularization (Ferdinand, Kirby, & Smith, 2019; Smith & Wonnacott, 2010), and technological innovation (Caldwell & Millen, 2008); and comparisons have been made of the effects that iterated learning has in different cultural domains (Tamariz, Kirby, & Carr, 2016). Iterated learning has also been studied in children (Flaherty & Kirby, 2008; Kempe, Gauvrit, & Forsyth, 2015; Raviv & Arnon, 2018) and other species (Claidière, Smith, Kirby, & Fagot, 2014; Fehér, Wang, Saar, Mitra, & Tchernichovski, 2009; Horner, Whiten, Flynn, & de Waal, 2006). For comprehensive reviews, see Scott-Phillips and Kirby (2010), Kirby, Griffiths, and Smith (2014), and Tamariz (2017). In brief, however, this body of research has broadly shown that iterated learning results in simple structures that are well adapted to the inductive biases of their learners. Of particular relevance to this thesis are studies of the iterated learning of semantic category systems (conceptual structures), which I return to in Section 2.1.3.

Iterated learning is sometimes thought of as a model of the historical process of cultural transmission and evolution – languages adapt to maximize their own transmissibility as they are passed from one generation to the next. But a proper understanding

of iterated learning recognizes the important role that *learning* plays in shaping the cultural behaviour being transmitted. Iterated learning is not simply a model of how languages are copied from one generation to the next; the important point is that learning is intimately involved in the process and that the mechanisms involved in learning are major contributing factors in moulding the structure of language. Ferdinand's (2015, p. 38) notion of *inductive evolution* is especially insightful in this regard:

**Inductive evolution:** The change over time of entities that replicate via a cognitive process of reverse engineering.

Every language user successfully acquires the language used in their speech community from an incomplete and impoverished dataset; an individual never observes every possible sentence but is able to produce novel sentences because they have reverse engineered the language's underlying grammar. This process of acquiring the language is accomplished by *induction*, which I define in the following way:

**Induction:** The inference of a general law from particular instances, supported by some form of prior knowledge or general heuristics.

I define induction this way to emphasize the fact that particular instances are not always sufficient to induce a general law that has good predictive power. Rather, the extraction of a general law from data is supported by 'prior knowledge or general heuristics', which I will refer to as the learner's *cognitive bias* or *prior bias*. In particular, this thesis emphasizes the idea that a rational learner can maximize the probability of correctly inferring how a language really works by applying Occam's razor as a heuristic; simpler explanations are, *ceteris paribus*, more likely to be true, a topic that we will return to in the following chapter. To be clear, I usually use the term 'induction' as shorthand for 'induction supported by a simplicity preference', but in Chapter 3 we will also consider a very different preference that learners might bring to the table.

That learners are biased towards simple explanations for the data they are confronted with has important ramifications for inductive evolution. Each naive learner gently nudges a language in the direction of greater simplicity, which often manifests itself in the form of greater structure, and when this process is repeated – iterated one generation after another – the effects of such simplification and restructuring accumu-



late, providing a potential explanation for why languages have some of the structural properties that they do.

With this in mind, one of the most important components of the iterated learning framework is what Brighton (2002) referred to as the *transmission bottleneck*:

**The transmission bottleneck:** A limit on how much information is transmitted from one generation to the next; the amount of data available to the learner from which to reverse engineer the language.

The bottleneck on transmission can also be thought of as Chomsky's (1980) *poverty of the stimulus*, although the emphasis is on the fact that the impoverished stimulus must be overcome not once, but generation after generation. One of the key concepts to understand about iterated learning is that the bottleneck controls how big the gaps are in the data, and therefore, the extent to which the learner must lean on their cognitive biases – whatever they might be – to fill those gaps in. Thus, under a tight bottleneck, where little data passes from one generation to the next, the cognitive biases that the learner brings to the table are very important in shaping the structure of the emergent language; in contrast, under a wide or fully open bottleneck, the learner's biases contribute little to its structure. This point becomes important in Chapter 3.

The bottleneck is, however, not the be-all and end-all. The bottleneck on transmission is one form of intergenerational information loss, which Spike, Stadler, Kirby, and Smith (2017) argue is an essential requirement in the emergence of structured languages. Other types of information loss – gaps in the data – that learners must respond to include a lack of exposure and various types of noise. These three types of information loss are formalized in the model presented in Chapter 3, and one of the findings shown in that chapter is that greater information loss through any of these means leads to simpler category systems under iterated learning.

## 1.4 Roadmap

Aside from its evolutionary approach to conceptual structure, this thesis interfaces with three other major areas of research in the language and cognitive sciences, and the goal of the following chapter, Chapter 2, is to provide some high-level background on these topics as they relate to the main content in Chapters 3 and 4:—

Chapter 2 first takes a brief look at the concepts and categorization literature, since, for the most part, I treat languages as collections of words/concepts/categories (terms that I largely treat as synonymous) that have structure in a meaning space à la Roger Shepard, Eleanor Rosch, and Peter Gärdenfors (among many others). Thus, I often talk about languages as *partitions* of a space, although technically, I would draw a distinction between a partition – the structuring of a meaning space into discrete categories (or concepts) – and a language – a labelling of such a partition.

Chapter 2 then looks at learning and how learners are able to reconstruct languages from limited, noisy data. This aspect of my work adopts Bayesian and information-theoretic approaches to inductive reasoning and emphasizes the role of simplicity as a fundamental bias that learners bring to the table, adopting ideas from Ray Solomonoff and Jorma Rissanen (among many others).

Finally, Chapter 2 turns to a relatively new strand of work on *informativeness* that is closely associated with Charles Kemp and Terry Regier (among many others). This work has argued that languages may be described as *informative* to the extent that they minimize the loss of information that occurs during communicative interaction; that is, informative languages minimize how much information is lost when an idea is transferred from one mind to another through the medium of language.

The thesis attempts to tie these strands of research together under one central claim:

**During the cultural evolution of languages and conceptual systems, inductive reasoning acts as a pressure for simplicity, while interactive communication acts as a pressure for informativeness. Combined, these two principal pressures give rise to languages that find an optimal balance between simplicity on the one hand and informativeness on the other.**

Chapter 3 deals with the simplicity-from-induction part of the claim and Chapter 4 deals with the informativeness-from-interaction part of the claim. Specifically, in Chapter 3, I formalize a Bayesian iterated learning model in which agents are instantiated with one of two prior biases – a bias for simplicity or a bias for informativeness; the predictions of this model are then tested in two experiments and we find that the bias for simplicity offers a much better account of human category learning. Then, in Chapter 4, I describe three iterated learning experiments – one with no pressure for informativeness, one with an artificial pressure, and one with a true pressure from a com-

---

municative task – and we show that more informative languages only arise when a true pressure from communicative interaction is present, leading to the emergence of higher level forms of linguistic structure. Chapter 5 is a short conclusion.



## Chapter 2

# Categorization, Compression, and Communication

The usual goal of communication is, of course, to set up “the same thought” in the receiver’s brain as is currently taking place in the sender’s brain. The mode by which such replication is attempted is essentially a drastic compression of the complex symbolic dance occurring in the sender’s brain into a temporal chain of sounds or a string of visual signs, which are then absorbed by the receiver’s brain, where, by something like the reverse of said compression—a process that I will here term “just adding water”—a new symbolic dance is launched in the second brain. The human brain at one end drains the water out to produce “powdered food for thought,” and the one at the other end adds the water back, to produce full-fledged food for thought.

— Douglas R. Hofstadter (2001)

In this quotation, Hofstadter vividly describes compression as it relates to the transmission of meaning from one mind to another. Under this view, a language can be thought of as a lookup table that converts meanings to signals, and also as a reverse lookup table that converts signals back to meanings; languages are the intermediaries through which communication occurs. However, languages are imperfect in this regard because they are lossy compressors: Once a thought has been compressed into a signal, it can never be fully reconstituted. Rehydrated food-powder is never as good as the original thing. We will return to this notion of compression towards the end of this chapter, the goal of which is to provide background material to the main content presented in Chapters 3

and 4. First, however, I set out a few findings from the concepts and categorization literature that will become relevant to this thesis later.

## 2.1 Concepts and Categorization

Because any object or situation experienced by an individual is unlikely to recur in exactly the same form and context, psychology's first general law should, I suggest, be a law of generalization.

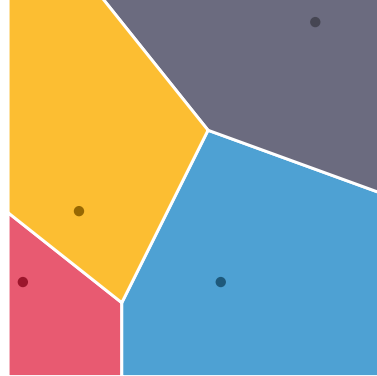
— Roger Shepard (1987)

We make sense of the world through a rich system of learned concepts, which allow us to categorize and make predictions about an infinite range of perceptual stimuli on the basis of their similarities to previous encounters. For example, although there are an infinite set of colours (or wavelengths of light), by categorizing them under a finite set of basic colour terms, we are able to represent, manipulate, and talk about colour in a useful way. Shepard (1987) referred to such concepts as 'consequential regions', since stimuli that are similar to each other – forming a compact region in the space of possible stimuli – are likely to have the same consequences for some organism; likewise, stimuli that have different consequences form separate consequential regions. As such, concepts tend to reflect the structure of the world, and they arise when the desire to generalize from one experience to the next conflicts with the simultaneous desire to distinguish between stimuli that differ in important ways.

### 2.1.1 Words, concepts, kinds, and categories

In this thesis, I take a very simplified view of language, reducing all its complexity to spaces that may be carved up into a small number of discrete categories. As a consequence of this dramatic simplification, the distinctions between 'word', 'concept', 'kind', and 'category' become somewhat blurred. Technically, I would draw the following distinctions: *natural kinds* exist out there in the world – there is some objective sense in which a flower is different from a bicycle; *concepts*, which usually approximate natural kinds, are the culture- or individual-specific mental representations we use to understand the world; and *words*, which symbolize concepts, are signals that particular languages use to permit communication.

**Figure 2.1:** Two-dimensional space discretized into four convex categories, yielding a Voronoi tessellation of the space. Each category is indicated by a different colour. The black dots indicate the ‘seeds’ or prototypes: All meanings (points) within a given category are closer to their associated prototype than to any other prototype. As such, the entire structure of the space can be represented minimally in terms of just the prototypes.



However, these distinctions break down in places, suggesting that the relationships between kinds, concepts, and words are more complicated. For example, the words a language provides affects performance on categorization tasks (e.g. Winawer et al., 2007), the mere presence of words helps us to make categorical distinctions (e.g. Sufill, Branigan, & Pickering, 2016), and the conceptual distinctions that languages make are influenced by the structure of the world (e.g. Perfors & Navarro, 2014). In this thesis, I treat words, concepts, and kinds as one and the same under the term ‘semantic category’ (or just ‘category’) because, for the purpose of this thesis, I have no particular interest in drawing a distinction between them. Essentially, this simplification amounts to assuming a one-to-one mapping between concept and word.

### 2.1.2 Convexity and compactness

Gärdenfors (2000) has argued that *convexity* is a fundamental property of concepts. Informally, a region of a metric space (i.e. a concept) is said to be convex if it is possible to travel in a straight line between any two points in that region without leaving it. More formally, a category  $C$  is said to be convex if, for any two points  $x$  and  $y$  in  $C$ , all points between  $x$  and  $y$  are also in  $C$ . A system of convex concepts that cover an entire space is known as a Voronoi tessellation, as depicted in Fig. 2.1. Convexity is especially interesting because it is naturally economical: An entire concept can be defined by a single point – the prototype; any novel meaning that is encountered can then be classified by finding the closest prototype. As such, the Voronoi tessellation and convex conceptual structures arise naturally from the fundamental operations involved in prototype-based categorization (Rosch, 1973). From this perspective, convex concepts may be considered equivalent to Shepard’s (1987) ‘consequential regions’.

Moreover, convexity appears to be a genuine property of at least some semantic domains; Jäger (2010), for example, has shown that the colour concepts documented in the *World Color Survey* (Kay, Berlin, Maffi, Merrifield, & Cook, 2009) are highly convex regions of colorimetric space. A question that remains open in the literature is why concepts tend to adopt such structure, and it seems to me that there are two main competing theories. First, the perspective from Shepard (1987) and Gärdenfors (2000), briefly alluded to above, says that convexity derives from the basic cognitive operations of categorization and generalization. Another example of work from this perspective comes from Steinert-Threlkeld and Szymanik (under review), who argue that ‘convexity can be explained by accounting for its role in the process of learning’. In his more recent work, Gärdenfors (2014) has adopted the view that pressure from communicative interaction might also be a contributing factor, citing computational simulations by Jäger and van Rooij (2007), who showed that convexity can arise in interaction because convex structures minimize the distance between intended and inferred meanings.

While convexity is an important concept, I do not take an especially strong position in this thesis on its status in natural languages and conceptual systems. Instead I adopt a weaker notion, which I call *compactness*:

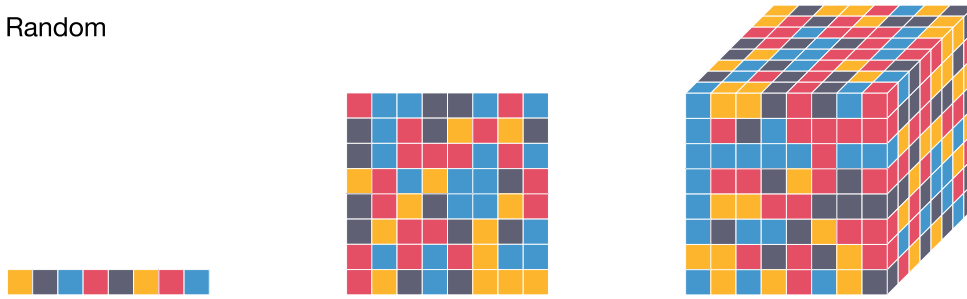
**Compactness:** The extent to which similar meanings belong to the same semantic category and dissimilar meanings belong to different semantic categories.

Similar notions already exist in the literature such as ‘family resemblance’ (Rosch & Mervis, 1975; Tversky, 1977) and ‘well-formedness’ (Regier, Kay, & Khetarpal, 2007); however, my use of the term ‘compactness’ is deliberately intended to be fluid and informal for the following reasons:

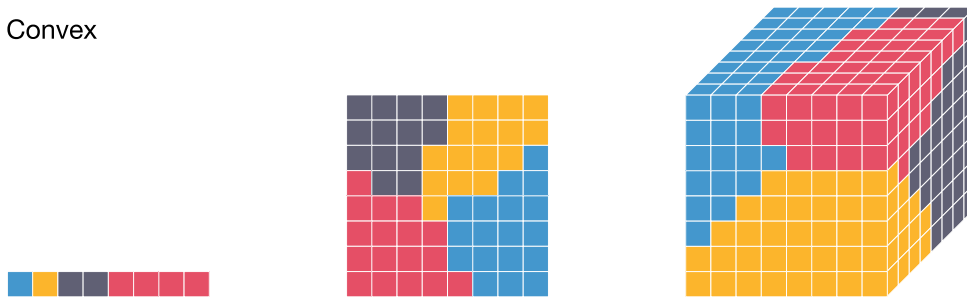
1. I do not want to make any specific claims about or limit myself to the stricter notion of convexity.
2. I do not want to imply *a-priori* that compact categories are ‘well-formed’ or ‘better’ than other arrangements (though I ultimately argue they are).
3. Compactness is intended to be a graded rather than binary notion.
4. The term retains a sense of similarity/distance but is general enough that I can apply it to structural features that were arrived at in very different ways.



Random



Convex



**Figure 2.2:** Illustrative examples of random (top) and convex (bottom) categorization systems in a one-dimensional, two-dimensional, and three-dimensional space. Each cell represents a meaning – a point in the space – and each colour represents a category – a set of meanings that are grouped together as one concept or labelled category. In these examples, the spaces are discretized into four categories. Under convex systems, the members of a category are tightly packed together, which is potentially beneficial to both learning and use.

By compactness, I simply mean that the categories in some partition are organized in such a way that similar (nearby) meanings are generally in the same category, while dissimilar (far apart) meanings are generally in different categories.

Nevertheless, although I adopt this flexible notion of compactness, partitioning a space into convex regions is computationally trivial,<sup>2</sup> so in the remainder of this chapter I will simulate convex partitions of a space to illustrate various points about the benefits that compactness confers. I will contrast these convex partitions with random partitions of the space, as illustrated in Fig. 2.2, and to simplify further, I will quantize the continuous space onto a grid of some discrete dimensionality.

<sup>2</sup> In a discrete space, one simply selects  $n$  meanings at random as ‘seeds’ – one seed for each of the target number of categories – and then classifies each of the remaining meanings according to the closest seed. In continuous spaces, a more complex approach is required, such as the Bowyer–Watson algorithm (Bowyer, 1981; Watson, 1981), which is the method I used to generate Fig. 2.1.

### 2.1.3 Iterated learning and conceptual structure

Having established a few preliminaries on concepts and categorization and returning to the iterated learning paradigm introduced in the previous chapter, I will briefly outline here four studies of the iterated learning of category structure in order to give a flavour of the previous work that this thesis builds on. We will return to these studies at various points in the thesis, but for a more complete review of this work, see Contreras Kallens, Dale, and Smaldino (2018).

J. Xu, Dowman, and Griffiths (2013) conducted one of the earliest examples of an iterated learning experiment using a continuous meaning space (as opposed to the discrete space adopted in Kirby et al., 2008). In their experiments, participants had to label a continuous colour space using between two and six colour terms according to condition. The way in which a participant discretized the space was then taught to a new participant in a chain. After 13 generations of cultural transmission, the structure of the space came to resemble the way in which colour space is typically structured by languages recorded in the *World Color Survey* (Kay et al., 2009). For example, in the three-term condition, the emergent systems discretized the space into dark, light, and red categories.

Perfors and Navarro (2014) used a meaning space of squares that vary continuously in terms of colour (white to black) and size (small to large). In one condition, there was an abrupt change in the *colour*, such that the stimuli could be categorized into two broad categories (light-coloured squares and dark-coloured squares); while in the other condition, there was an abrupt change in the *size* of the squares. Labels for these stimuli were then passed along a transmission chain of learners. In both conditions, the authors found that the structure of the emergent languages came to mirror the structure of the meaning space, primarily making colour or size distinctions according to condition.

Silvey, Kirby, and Smith (2013) produced a continuous meaning space by randomly generating four seed polygons and then gradually morphing the polygons into each other, creating a space of 25 stimuli. The space had no obvious internal boundaries; as such, participants showed variation in how they discretized it. The authors also conducted an iterated learning experiment using the same meaning space (Silvey, 2014, Chapter 5); (Silvey, Kirby, & Smith, 2015). In this experiment, each generation consisted of a pair of participants who communicated about the stimuli using a fixed set

of up to 30 words. Over five generations, the category systems that emerged tended to make fewer distinctions and became easier to learn. Furthermore, the category structures became increasingly convex.

Of particular relevance to the next chapter is an iterated learning experiment by Canini, Griffiths, Vanpaemel, and Kalish (2014). The authors looked at how four different concept structures – varying from easy to hard – are learned in four different stimulus spaces – varying in the separability of the dimensions (i.e. the extent to which the dimensions of the space can easily be individuated). Their results replicated a wealth of findings from decades of concept learning research, suggesting that the iterated learning paradigm can be used to reveal human inductive biases – especially the bias for simplicity. For example, they found that concept structures marking a distinction on one dimension are easier to learn than those marking distinctions on two dimensions, especially when the stimuli are separable (replicating findings by e.g. Ashby & Maddox, 1990; Goudbeek, Swingley, & Smits, 2009; Shepard, Hovland, & Jenkins, 1961). They also found that people can switch between rule-based, prototype-based, and exemplar-based modes of learning depending on the type of concept structure to be learned (as argued by e.g. Ashby & Maddox, 2005; Erickson & Kruschke, 2002; Nosofsky, Palmeri, & Mckinley, 1994).

## 2.2 Simplicity and Learning

Objectivity is reached when we have squeezed out every last bit of information from the data until nothing but random noise remains.

— Jorma Rissanen (1989)

Simplicity is argued to be a fundamental principle of cognition (Chater, Clark, Goldsmith, & Perfors, 2015; Chater & Vitányi, 2003; Feldman, 2016) that can explain the kinds of structure we find in languages (Culbertson & Kirby, 2016) and conceptual systems (Kemp, 2012). The application of a simplicity principle to problems of inductive reasoning is usually attributed to William of Ockham, whose maxim – Occam’s razor – states that, *ceteris paribus*, simple hypotheses should be preferred over complex ones. Occam’s razor is often contrasted with the Epicurean principle of multiple explanations, which states that all hypotheses consistent with observations should be retained,

regardless of how complex they are. This section reviews several perspectives on the simplicity principle and its relationship with learning.

### 2.2.1 Bayesian inference as a model of learning

Ultimately, a rational learner wishes to gain an accurate understanding of some phenomenon that exists out in the world – accurate enough that the learner is able to use that understanding to make predictions about how the phenomenon will behave in the future. In other words, the learner wishes to gain knowledge about a system that has good predictive power. However, the learner must attempt to induce this knowledge in the face of incomplete, noisy data. Bayesian inference provides a rational model for doing exactly this. Given data  $d$ , the learner ought to select the hypothesis  $h$  that maximizes the posterior probability  $p(h|d)$ , which, by Bayes' theorem, is given by:

$$p(h|d) \propto p(d|h)p(h). \quad (2.1)$$

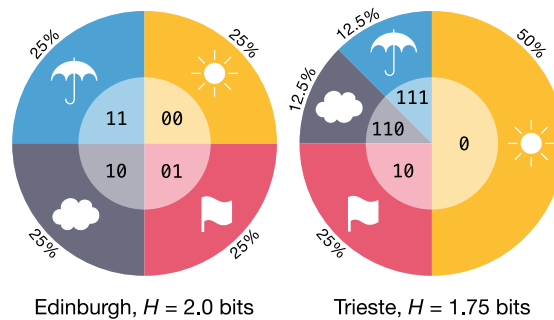
In words: The probability that hypothesis  $h$  is true (given that we have observed data  $d$ ) is proportional to the likelihood of observing  $d$  (assuming  $h$  were true) multiplied by the probability of  $h$  according to our prior expectations.

According to this formalization, learning (induction) may be viewed as weighing up two pieces of evidence. The likelihood term,  $p(d|h)$ , captures how well the data supports the hypothesis, while the prior term,  $p(h)$ , captures any prior evidence (innate knowledge, for example) that the learner brings to the table. This thesis, and in particular the content of Chapter 3, starts from the following observation:

**The rational learner, who has no prior expectations in some particular domain, should at least apply Occam's razor because simpler explanations are inherently more probable.**

This insight was first rigorously formalized by Solomonoff (1964a, 1964b), who was seeking a principled way to set the prior in problems of Bayesian induction. As we shall see later in this section, when coupled with a simplicity-based prior, Bayes' rule becomes a mathematical formalization of Occam's razor *and* the principle of multiple explanations: The likelihood retains hypotheses consistent with the data, while the prior places greater weight on hypotheses that are simple and therefore more likely to be true

**Figure 2.3:** Pie charts representing the probability of four weather conditions in Edinburgh and Trieste, along with their corresponding codewords in an optimal prefix-free code. When the weather is more predictable, as in Trieste, shorter codewords may be assigned to frequent states and longer codewords to infrequent states, resulting in shorter average codeword length (i.e. entropy  $H$ ).



in the absence of any other information (Li & Vitányi, 2008, p. 347). An ideal model of such a learner therefore requires some principled way to calculate the *a priori* simplicity of a hypothesis. In this thesis, I adopt the view that the extent to which a language hypothesis is compressible is a practical estimator of its simplicity. To understand why, I will first lay some groundwork and then describe the various ways in which compression has been deployed in the field.

## 2.2.2 Coding and probability: Some preliminaries

Let's imagine that there are four possible weather conditions – sunny, windy, cloudy, and rainy – and that we would like to transmit the current weather conditions between two cities, Edinburgh and Trieste. The communication channel is very costly to use, and as such we would like to transmit the weather reports in as few bits as possible. We know in advance that the weather in Edinburgh is very unpredictable: at any given moment there is a 25% chance that the weather might be either sunny, windy, cloudy, or rainy. Therefore, in designing the binary code that we will use to transmit the weather reports, we opt for a system in which each weather state is represented by a unique two-bit codeword: 00 for sunny, 01 for windy, 10 for cloudy, and 11 for rainy, as shown in Fig. 2.3. Therefore, every transmission involves sending 2 bits of information. In Trieste, by comparison, the weather is more predictable – it tends to be sunny more frequently, while rain and cloud are less frequent. Therefore, to minimize the cost of using the communication channel, we opt for a coding system in which sunny weather is represented by a single bit, say 0, while the other states are represented by two or three bit codewords, as highlighted in Fig. 2.3. On average, each transmission requires just 1.75 bits, since the one-bit codeword will be used very often (50% of the time), while each of the three-bit codewords will only be used rarely (12.5% of the time).

The coding systems we have devised here are no accident; each is optimal for its respective city. Given that we know the probability with which a state  $x$  occurs – for example, that it is sunny 50% of the time – the shortest possible binary codeword that may be devised to represent that state has a length of  $-\log p(x)$  bits.<sup>3</sup> This, however, does not tell us what that codeword should be and a coding algorithm, such as Shannon–Fano, Huffman, or arithmetic coding, must be used to derive a set of codewords that get close to the Shannon entropy (Shannon, 1948), which is, in fact, a formalization of the intuition given in the previous paragraph; the entropy of a set of possible states  $X$  is the probability of state  $x$  occurring multiplied by the length of an optimal codeword used to represent that state, summed over all states:

$$H(X) = \sum_{x \in X} p(x) \cdot -\log p(x), \quad (2.2)$$

and it represents a lower bound on the expected length of a codeword transmitted over the channel. The lower the entropy, the more predictable the system.

Optimal codes may be designed using variable-length, prefix-free codewords, as is the case in our example above. By prefix-free, we mean that no codeword is a prefix of any other codeword and, therefore, the state represented by a codeword is entirely unambiguous, despite the fact that codewords have differing lengths. On reception of a 0 from the Trieste weather station, we know that the weather must be sunny, but if we receive a 1, we need to keep listening; if the next bit is a 0 it must be windy, but if it is a 1, we need to keep listening; the third bit will then allow us to disambiguate between cloudy and rainy. Although couched in coding theory, prefix-free codes are, in fact, just a different way of thinking about probabilities; moreover, if we abstract away from the idea that a code has to be an actual string of binary digits and allow for noninteger codelengths, we find that codelengths and probabilities are, underlyingly, the same mathematical construct. If we know the probability of some state, we can transform that probability into a codelength:

$$\text{codelength}(x) = -\log p(x). \quad (2.3)$$

---

<sup>3</sup> All logarithms in this thesis are to base 2.

But, equally, if we know the codelength of a state, we can find out its probability:

$$p(x) = 2^{-\text{codelength}(x)}. \quad (2.4)$$

As we shall see shortly, thinking about probabilities in terms of codelengths offers useful insights into rational Bayesian induction, especially in the sense that knowing the most compressed codeword – or shortest description – of a hypothesis tells us something about the prior probability of that hypothesis, and therefore the extent to which our rational, Occam’s-razor applying learner should favour that hypothesis *a priori*.

### 2.2.3 Sources of compression in language structure

The compression of data refers to the encoding of that data in such a way that it requires fewer bits to store or transmit. As we saw in the previous section, knowing the probability of some state allows us to derive the shortest possible unambiguous codeword for that state. However, this alone is not compression. When we seek to compress a dataset, we can also take advantage of the regular patterns and structures that it contains. For example, if the weather report from Edinburgh at time  $t$  is cloudy, then there might be a greater than 25% probability that it will be raining at time  $t + 1$ . Similarly, there may exist higher levels of structure, such as seasonal trends in weather patterns that might offer further opportunities for compression. The difficulty for a compressor, then, is in identifying the structure in some dataset and in deciding which level of structure it is best to take advantage of. Naturally, compression has been very widely studied in computer science, leading to various algorithms, such as Lempel–Ziv–Welch (Welch, 1984; Ziv & Lempel, 1978), that seek to find shorter representations of data.

I can think of at least three ways in which compression is often discussed in the context of language, and at this point it will be useful to enumerate these so that I can position this thesis in the broader context.

**Compression from communication** In the quotation from Hofstadter (2001) that opened this chapter, compression was related to language in terms of how signals are structured for efficient transmission between interlocutors. In this sense, a language may be thought of as a tool for compressing a complex meaning into a short signal that may be rapidly transmitted to a listener, who decompresses the signal on the other

end. Various strands of research have found, for example, that languages adopt short strings for frequently used meanings and long strings for infrequently used ones and that natural languages tend to be close to optimal in this regard (see e.g. Ferrer i Cancho & Solé, 2003; Kanwal, Smith, Culbertson, & Kirby, 2017; Piantadosi, Tily, & Gibson, 2011, for various perspectives on this). Broadly, this view posits that principles of effort minimization in communicative interaction act as sources of compression in language structure. For the most part, this is *not* the sense in which I use compression in this thesis, although the model of communication discussed in Section 2.3.2 does indeed take for granted that signals are optimized in this way.

**Compression from learning I: Cognitive economy** The second way in which compression is relevant to language is in terms of cognitive economy or how easy it is to learn and store a language. From this perspective, compressed language structure derives from the imposition of cognitive constraints during the learning and storage of a language. For example, in Section 2.1.2, we saw how systems of convex concepts are naturally very economical, since each concept can be represented by a single point (the prototype). Gärdenfors’s (2000) motivation for the naturalness of convexity in conceptual structure is rooted in cognitive economy: ‘I believe that [convexity] can be defended by a principle of *cognitive economy*; handling convex sets puts less strain on learning, on your memory, and on your processing capacities than working with arbitrarily shaped regions’ (p. 70). Indeed, this view has recently been supported by Sims (2018), who shows how efficient coding of perceptual stimuli gives rise to Shepard’s (1987) universal law of generalization.

Kemp and Regier’s (2012) work on kinship categorization systems, which will be discussed in a lot more detail later, also takes the view that categorization systems are simple (compressible) ‘to the extent that [they] can be concisely mentally represented and therefore easily learned and remembered’ (p. 1049), suggesting that the authors view compressible conceptual systems as deriving from issues of cognitive economy in learning and memory. The idea that compressed structures derive from issues of cognitive economy has also been explored in many other contexts, including, for example, chunking behaviour (Mathy & Feldman, 2012), language acquisition (Wolff, 1982), and working memory (Chekaf, Gauvrit, Guida, & Mathy, 2018). Broadly, this view posits that principles of efficient coding in learning act as sources of compression in language.



**Compression from learning II: Inductive reasoning** Finally, and as mentioned in passing above, compression is relevant to the *induction* of language – that is to say, it is relevant when an agent weighs up hypotheses in terms of their fit to the data (the likelihood) and the extent to which they satisfy Occam’s razor (which may be represented in the prior). This is because the compressibility of a candidate language hypothesis is an estimator of its simplicity. Culbertson and Kirby (2016), for example, have emphasized this perspective in the context of language learning, arguing that a domain-general simplicity principle has domain-specific effects in language structure.

Communication acts as a pressure for compressibility in the sense that short, compressed signals are more efficient to transmit, although this is not the sense adopted in this thesis. Rather, the view I take herein is that the pressure for compressibility derives from learning, and, as we have seen, this may be motivated in at least two ways. First, the process of mentally encoding the structure of the language will tend to result in simpler languages; from this perspective the brain seeks compressed representations due to memory or processing constraints. Second, the process of inducing the structure of a language will also tend to result in simpler languages; from this perspective the brain seeks compressible explanations because they are, by Occam’s razor, more likely to be true. I do not take a particularly strong view on which of these – compressed representations or compressible explanations – is ‘correct’; in fact, it may well be the case that both contribute to compressibility in language and category structure. That being said, I tend to prefer the ‘compressible explanations’ idea because, as we shall see, it has been tightly formalized, and it is not dependent on any particular theory about how the brain encodes information; even if the brain had infinite processing power and memory, it would still seek compressible explanations.

In the following section, we will formalize the relationship between compression and inductive reasoning, and we will see how compression offers a natural extension of the Bayesian framework introduced at the start of this chapter.

#### 2.2.4 Algorithmic probability and the minimum description length principle

The length of the shortest computer program that generates a given string is known as the string’s Kolmogorov complexity. A string that is generated by a short program

has low complexity, and a string that is generated by a long program has high complexity. Kolmogorov complexity was discovered independently by Chaitin (1969), Kolmogorov (1965), and Solomonoff (1964a, 1964b). Solomonoff's central insight was to think of observed phenomena as outputs from programs running on a universal Turing machine. Given some input (a program), the machine produces some output (a phenomenon). The probability of some phenomenon occurring is therefore the combined probability of all the programs that produce that phenomenon, all ways in which that phenomenon could have been generated. Simpler phenomena are generated by more programs with a shorter length and are therefore *a priori* more probable events in the universe. Solomonoff referred to this as the 'universal prior' (e.g. Solomonoff, 1997, p. 83), since the only universal way to compute the prior probability of some hypothesis is to consider the infinite possible ways in which that hypothesis could be justified. As such, the universal prior, like Kolmogorov complexity in general, is incomputable.

Inspired by Solomonoff's work and applying his ideas to concrete problems, Rissanen (1978) formulated the minimum description length (MDL) principle. Essentially, MDL-based methods fix a description code in which hypotheses can be expressed, and the universal prior is approximated by the length of that description. The MDL principle is often used to address a central concern in practical model selection problems: The selection of an overly complex model that, while fitting the observations well, predicts future observations poorly. MDL guards against such overfitting by requiring not only that a model fits the observations well, but also that the model is sufficiently simple, determined by how compressible that model is.

By Bayes' theorem (Equation 2.1), we may calculate a posterior probability distribution over some hypothesis space,  $H$ , from which an idealized learner is expected to select the hypothesis,  $h_{\max}$ , that has the greatest probability of being true given the data:

$$h_{\max} = \arg \max_{h \in H} [p(h|d)] = \arg \max_{h \in H} [p(d|h)p(h)]. \quad (2.5)$$

Treating the probabilities as codelengths yields an equivalent minimization problem:

$$h_{\max} = \arg \min_{h \in H} [-\log p(h|d)] = \arg \min_{h \in H} [-\log p(d|h) - \log p(h)], \quad (2.6)$$

which may be rewritten in terms of description lengths, DL, as:

$$h_{\max} = \arg \min_{h \in H} [\text{DL}(h|d)] = \arg \min_{h \in H} [\text{DL}(d|h) + \text{DL}(h)]. \quad (2.7)$$

This tells us that the most probable hypothesis is one which minimizes the sum of two codelengths or, in other words, the length of two descriptions: A description of the data given the hypothesis,  $\text{DL}(d|h)$ , and a description of the hypothesis itself,  $\text{DL}(h)$ .

What we have derived here is the MDL principle (Grünwald, 2007; Rissanen, 1978). The key insight behind MDL is that treating probability as description provides us with a concrete way to determine how probable some hypothesis is. For example, imagine we observe the following 264-bit sequence:

```
010011100110000101110100011101010111001001100101001000000110100101
110011001000000111000001101100011001010110000101110011001001110110
010000100000011101110110100101110100011010000010000001110011011010
010110110101110000011011000110100101100011011010010111010001111001
```

The sequence appears random and contains no obvious structure. One hypothesis we might have is that the generating process hardcodes this sequence verbatim; formulated as a Python program,<sup>4</sup> our hypothesis about the generating process might look like this:

```
print('01001110011000010111010001110101011100100110010100100000011
010010111001100100000011100000110110001100101011000010111001100100
111011001000010000001110111011010010111010001101000001000000111001
10110100101101101111000001101100011010010110001101101001011101000
1111001')
```

This description of our hypothesis is quite long,  $\text{DL}(h) = 273$  characters, but the description length of the data given the hypothesis,  $\text{DL}(d|h)$ , would be zero; it would be unnecessary to describe the data itself because the hypothesis already generates it exactly. Alternatively, we might hypothesize that the generating process is random:

```
from random import randint
print(''.join([str(randint(0,1)) for _ in range(264)]))
```

<sup>4</sup> Although I formulate hypotheses as Python programs here, note that MDL methods typically fix a much more restricted code in which data and hypotheses can be represented, an example of which will be described in the next section. Of course, the description lengths we compute will depend on our somewhat arbitrary choice of description code, but that is the price we pay for making the incomputable computable.

While this hypothesis is comparatively short,  $DL(h) = 82$  characters, it almost certainly will not generate the observed sequence, so the description length of the data given the hypothesis,  $DL(d|h)$ , will have to be quite long to list all the errors the program makes in reproducing the observed data.<sup>5</sup>

After some thought, we might hypothesize that the sequence is actually an English sentence encoded in 8-bit ASCII:

```
s='Nature is pleased with simplicity'
print(''.join(['{0:b}'.format(ord(c)).zfill(8) for c in s]))
```

This hypothesis is also quite short,  $DL(h) = 98$  characters, but it fits the data perfectly. As such, this hypothesis minimizes the sum of  $DL(h)$  and  $DL(d|h)$ , making it the most probable hypothesis about the process that generated our observed sequence. This captures the intuition that a good hypothesis is one that predicts the observed data well but is also concise, making as few assumptions as possible about the true process.

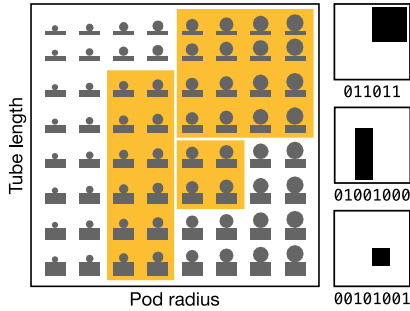
This brings us to the quotation from Rissanen (1989, p. 88) that opened this section: ‘objectivity is reached when we have squeezed out every last bit of information from the data until nothing but random noise remains’. By ‘objectivity’ here, Rissanen means something like ‘getting as close as possible to the objective truth given the data available to us’. In the act of compressing some observed data – finding a hypothesis that optimizes the tradeoff between concision and predictive power – we ‘learn’ something about the process that generated that data, allowing us to draw a formal equivalence between learning and compression (Grünwald, 2007, p. 91).

### 2.2.5 Complexity and concept learning

The classic work on concepts and categorization showed that certain types of concept are harder to learn than others – that concepts on many dimensions are harder to learn than concepts on just one, for example (e.g. Shepard et al., 1961). However, this body of work largely left the *why* question unanswered – why should one conceptual structure be harder to learn than another? Feldman’s (2000) answer to this relates concept learning to compressibility, providing empirical evidence for the ideas described above. In

<sup>5</sup> If our hypothesis proposes that the data was generated randomly, then  $p(d|h) = 1/2^{264}$  or, in other words,  $\text{codelength}(d|h) = -\log 1/2^{264} = 264$  bits. Therefore, the hypothesis that we have a random process will have approximately the same overall description length as the hypothesis that we have a deterministic process that generates this particular 264-bit sequence.

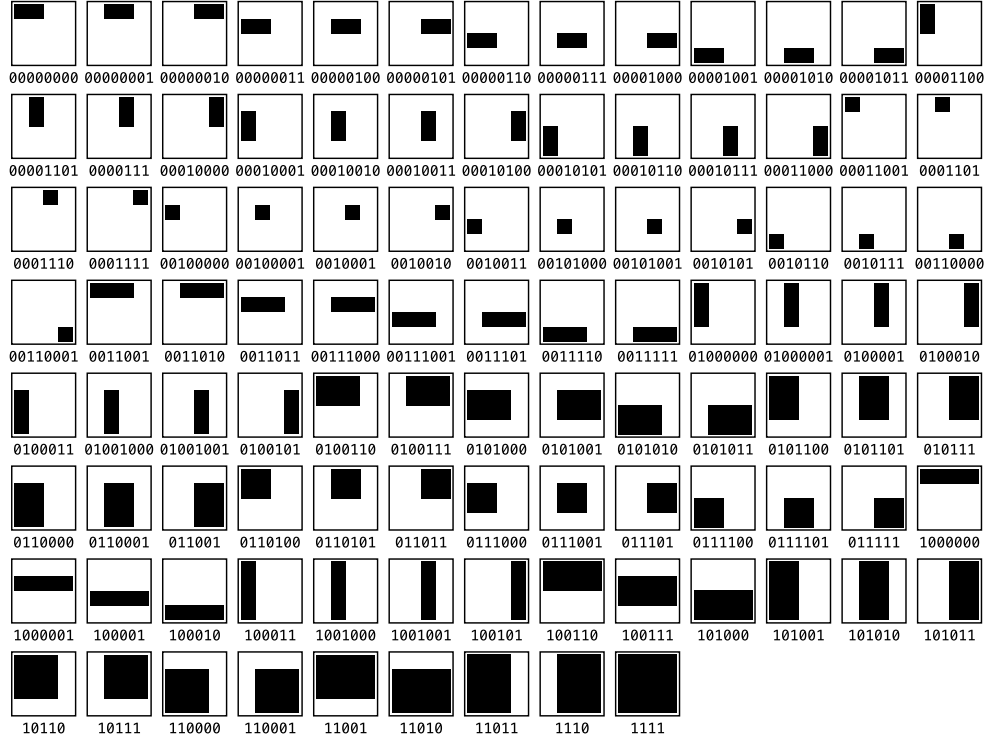
**Figure 2.4:** The stimulus space used in Fass and Feldman (2002). The spacecraft vary continuously on two dimensions: the length of the ‘tube’ and the radius of the ‘pod’. Participants had to learn to identify enemy spacecraft, the ones highlighted in yellow, for example. The complexity of a concept was estimated by its description length in the rectangle code (see main text and Fig. 2.5); in this case, the concept can be described using three rectangle symbols, yielding the binary description 0110110100100000101001 (22 bits).



this work, Feldman (2000) showed that participants’ performance on a range of different conceptual structures has a logarithmic relationship with their Boolean complexity; concepts that have longer logical descriptions are harder to learn and *vice versa*.

Extending this to continuous concepts, Fass and Feldman (2002) made use of the MDL principle to formulate a model of concept induction and they compared the predictions of this model to a concept learning experiment. In the experiment, participants had to learn how to classify ally and enemy spacecraft from a two-dimensional continuous space of spacecraft stimuli; the stimuli consisted of two parts, a tube and a pod, which varied in size as shown in Fig. 2.4. Participants completed 12 rounds, each of which tested a particular partition of the space (i.e. a particular conceptual structure). In each round, a participant played a five-minute game in which they had to destroy enemy ships and allow ally ships to land. The authors showed that participants’ performance on a given partition was correlated with its minimum description length; participants performed better on partitions that had a shorter two-part description (i.e. the description length of the data given the hypothesis plus the description length of the hypothesis). This, they argue, provides evidence that MDL-based methods offer a good account of human concept learning.

Fass and Feldman (2002) referred to their description method as the ‘rectangle language’, and their method is used extensively in this thesis to measure the complexity of languages; as such, I refer to their method as the rectangle *code* to avoid confusion. The rectangle code consists of a set of rectangle symbols – as shown in Fig. 2.5 – which may be used to describe arbitrary regions (i.e. concepts) in the meaning space, an example of which was highlighted in Fig. 2.4. The details of this method are discussed in a lot more detail in Paper 2 and in Section 3.4. For now, however, it suffices to say that the complexity of a category (a region in the space), and therefore a language or category system



**Figure 2.5:** The 100-symbol alphabet used to describe categories in a 4×4 space. For illustrative purposes, the symbols are shown with binary codewords: Probable symbols (i.e. rectangles) are assigned short codewords and improbable ones are assigned long codewords.

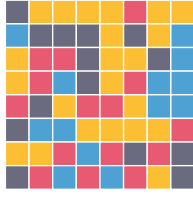
as a whole, can be estimated according to its shortest description in the rectangle code. Other complexity measures, based on the same compressibility principles, have also been reported in the literature, such as the block decomposition method (Zenil, Soler-Toscano, Delahaye, & Gauvrit, 2015; Zenil, Soler-Toscano, Dingle, & Louis, 2014), and, of course, techniques from image compression may also provide useful insights, such as chain code compression (Freeman, 1961).

## 2.2.6 Hallmarks of simple category systems

If we adopt the rectangle code as a measure of how compressible a language is, what kinds of language are predicted to be simple? To answer this, I have generated category systems in an 8×8 space<sup>6</sup> by assigning each of the 64 meanings to one of  $n$  categories, where  $n$  may vary from 1 to 64. When  $n = 1$ , all meanings belong to a single category

<sup>6</sup> Note that I have switched here to using an 8×8 version of the rectangle code, not the 4×4 one illustrated in Fig. 2.5. For the most part, I use a 4×4 space for illustrative examples and an 8×8 space for actual results.

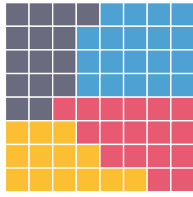
Four category language with a random structure



```
00100100000100000000111100100110010001001101100010
01110001000000011111001010001100010100100010010100
10100010100110000101001111001010100001001010100100
01000111000000011110010000010011000010010100000010
01100000001001110010010100000000010100101100000111
00000010100111000100100011000001001000010000111100
10011001100000010010000101001001001010100000001010
1000100010101011000010111000001001011110000000011
00010010100010000010010010010011101000100111101000
1011111000000010101000101010011
```

complexity  $\approx$  481 bits

Four category language with a convex structure



```
01001001001001010000100100010110000010101010000111
10001100011001000000100011111100010000100001011011
011010010100101001001000001010011
```

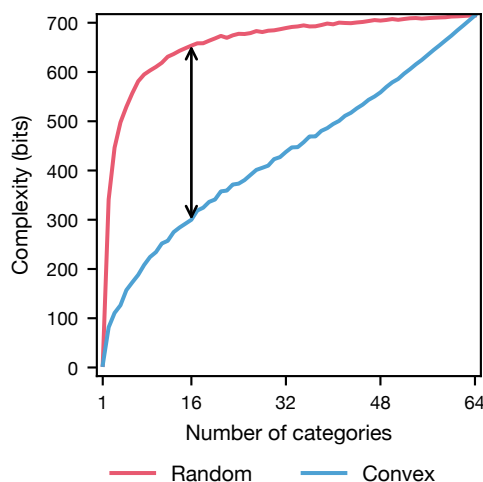
complexity  $\approx$  132 bits

**Figure 2.6:** Complexity of languages that have random vs. convex category structure in the rectangle code. Categories that have a random structure have a long description length and are therefore complex; categories that have a more compact structure have a comparatively short description length and are therefore simpler.

(the ‘trivial partition’), and when  $n = 64$ , each meaning forms its own distinct category (the ‘partition of singletons’). The trivial partition and partition of singletons represent the two extremes, the simplest and most complex systems respectively. In addition, we will compare two classes of category system – random and convex. In random systems, the meanings are assigned to categories entirely randomly; in convex systems, the space is partitioned into  $n$  randomly-generated, convex categories. Illustrations of each of these types of system are shown in Fig. 2.6 for  $n = 4$  categories.

Fig. 2.7 plots the complexity of these two classes of system as a function of  $n$ . These results demonstrate that there are primarily two ways in which a category system may be considered simple:

1. **Category sparsity:** The fewer categories there are, the simpler the system is as a whole; a system that makes few categorical distinctions is simple because less information is required to draw those distinctions.
2. **Compactness:** The more compact categories are, the simpler the system is as a whole; compact categories are simple because groups of clustered meanings may be described more succinctly in terms of higher level structure.



**Figure 2.7:** Complexity of simulated category systems in an 8×8 space. The pink curve shows the mean complexity of 100 randomly generated category systems as a function of the number of categories. The more categories a system has, the more complex it is. The blue curve shows equivalent results for *convex* category systems. Systems that consist of compact categories, such as the convex categories generated here, are much simpler than unstructured systems. This gives rise to a compactness advantage, illustrated by the black arrow: For a given number of categories (e.g. 16), a compact system is simpler than a random system.

Systems that consist of few categories are simpler than those that consist of many, and, for a given number of categories, systems that consist of compact categories are simpler than those that consist of noncompact categories. This gives rise to a *compactness advantage*, illustrated in Fig. 2.7 by the black arrow, an advantage that we will return to later in the chapter.

In this section, we have seen how language learners are expected to seek the simplest explanation of the observed data. In iterated learning, the effects of this bias for simplicity are amplified. Each new language learner – faced with the task of inducing a probable hypothesis from an impoverished dataset – moves the language a little bit closer to their prior bias, such that in the limit, iterated learning converges on a distribution of languages that reflects whatever that prior bias might be (Griffiths & Kalish, 2007). This is interesting for two reasons. Firstly, from a theoretical perspective, it suggests that learning – specifically, induction under a simplicity bias – will have a simplifying effect on a language in the long term, providing an explanation for why languages tend to possess simple, regular, compressible patterns (this is the perspective taken by e.g. Kirby et al., 2015, see Chapter 1). Secondly, from a methodological perspective, the iterated learning paradigm offers a useful means for revealing what language learners’ inductive biases actually are (this is the perspective taken by e.g. Canini et al., 2014, see Section 2.1.3). We will now put simplicity to one side to consider a rather different aspect of language: its use in communicative scenarios.



## 2.3 Informativeness and Communication

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point.

— Claude Shannon (1948)

Shannon and Weaver (1949, p. 4) set out three levels at which information may be lost in the process of communication:

- A. How accurately can the symbols of communication be transmitted? (The technical problem)
- B. How precisely do the transmitted symbols convey the desired meaning? (The semantic problem)
- C. How effectively does the received meaning affect conduct in the desired way? (The effectiveness problem)

This thesis largely abstracts away from Level A, which involves issues of noisy channels, error-correcting codes, redundancy, and so forth (for some perspectives on this see e.g. Plotkin & Nowak, 2000; Winter & Wedel, 2016). Instead, the notions of informativeness and communication discussed in this section exist firmly at Level B: Languages are informative to the extent that meaning is conveyed with high precision during the communicative process. Later in this section, we will see how this notion of informativeness may be formalized as *communicative cost*, a measure of how much information is lost, on average, when the listener tries to reconstruct the speakers intended meaning through the medium of some particular language.

Whereas the problem at Level B is about ensuring the listener acquires the same meaning as the speaker, the problem at Level C is concerned with what the listener does with that information in a given communicative scenario. In general, I refer to this distinct notion as *communicative accuracy* – the extent to which the speaker’s desired result actually occurs following the communication of some thought. Put another way, the informativeness of a language is measured by its communicative cost and languages may be costly to use or noncostly to use depending on the particular set of design choices they follow. This issue of informativeness exists at Level B. In contrast, Level C

is concerned with the outcome of a particular communicative interaction, in which a pair of interlocutors may be successful or unsuccessful (or indeed successful to a certain degree) depending not only the particular language they happen to be using, but also the contextual knowledge that speaker and listener bring to the table and other such external factors.<sup>7</sup> Of course, success at each level is in part determined by success at the level below. The accurate transmission of signals (Level A) contributes to how precisely meaning is recovered by the listener (Level B), which in turn contributes to the probability of a successful outcome (Level C).

In the remainder of this section, we will first look at some examples of how languages have been studied in terms of informativeness, and we will then turn our attention to the more technical issue of communicative cost. I then describe the hallmark features of informative category systems and contrast them with the hallmark features of simplicity that we identified at the end of the last section.

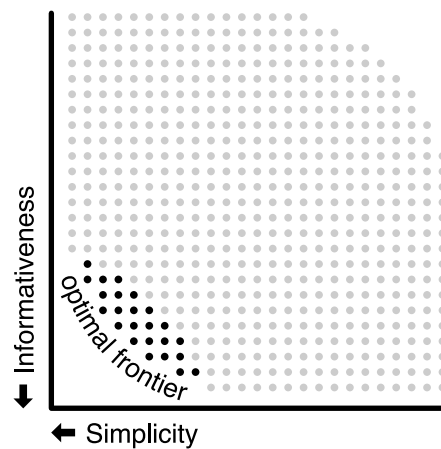
### 2.3.1 Informativeness in typological datasets

In the first work of its kind, Kemp and Regier (2012) conducted a study of kinship terms using a large dataset originally collected by Murdock (1970). In this work, the simplicity of a kinship system was measured following the same principles described in Section 2.2: A kinship system is simple to the extent that it has a short description in a code for representing family members. In English, for example, the concept *MOTHER* is defined by the logical expression  $\text{mother}(x, y) \leftrightarrow \text{parent}(x, y) \wedge \text{female}(x)$ ;  $x$  is a mother to  $y$  if  $x$  is a parent to  $y$  and  $x$  is female. This measure of simplicity was then paired with their measure of informativeness, communicative cost, which is explained in detail shortly, and they showed that, of the approximately  $10^{55}$  kinship systems that could exist (i.e. possible partitions of a family tree), only a tiny fraction were attested in natural languages. More importantly, they showed that natural languages exist at what they call the ‘optimal frontier’ of simplicity and informativeness: Kinship systems are both maximally simple and maximally informative (see Fig. 2.8 for an illustration).

Kinship terms are generally discrete; an individual may be a son or a daughter with-

<sup>7</sup> I take care to point this distinction out because there are points in the thesis, especially, in Chapter 3, where it may not be immediately obvious which level is being discussed, since communicative cost may be formulated with a distance metric, but equally communicative accuracy may be quantified in terms of distance between intended and inferred meanings.

**Figure 2.8:** Each point represents a possible language in simplicity–informativeness space. The optimal frontier represents the best tradeoff that may be achieved between the two properties. Typological studies in various domains show that real languages cluster around the optimal frontier (black dots), suggesting that languages are adapted to be optimally simple and optimally informative.



out gradations in between. Other domains, however, are naturally more continuous, and Regier and colleagues have extended their methods to such domains. One example of this is an analysis of colour terms by Regier, Kemp, and Kay (2015, Case Study 1), building on work by Regier et al. (2007) and using data from the World Color Survey (Kay et al., 2009). Regier et al. (2015) extended their notion of communicative cost to continuous spaces by integrating a measure of perceptual similarity. Like the kinship study described above, they compared naturally-occurring partitions of colour space to possible partitions that hypothetically could exist, and they showed that extant colour systems are more informative than would be expected by chance.

To give one final example, Y. Xu, Regier, and Malt (2016) were interested in the idea that historical processes of word derivation (such as analogy and metaphor) result in languages that are nonoptimally informative because, over historical time, a word will come to describe a wide variety of referents that are not necessarily similar to each other. They refer to this as historical semantic chaining. For example, the word *port*, originally meaning gateway or opening, was later extended to physical computer inputs and then to virtual channels used in networking. Thus, the word *port* could be considered uninformative because – pragmatics aside – it does not carry information about which particular meaning is meant. They tested this with words for household containers in three languages (using data from Malt, Sloman, Gennari, Shi, & Wang, 1999) and found that, although the words were consistent with semantic chaining, the languages were nevertheless more informative than would be expected by chance.

The authors, Regier in particular, have taken the same general approach in a variety

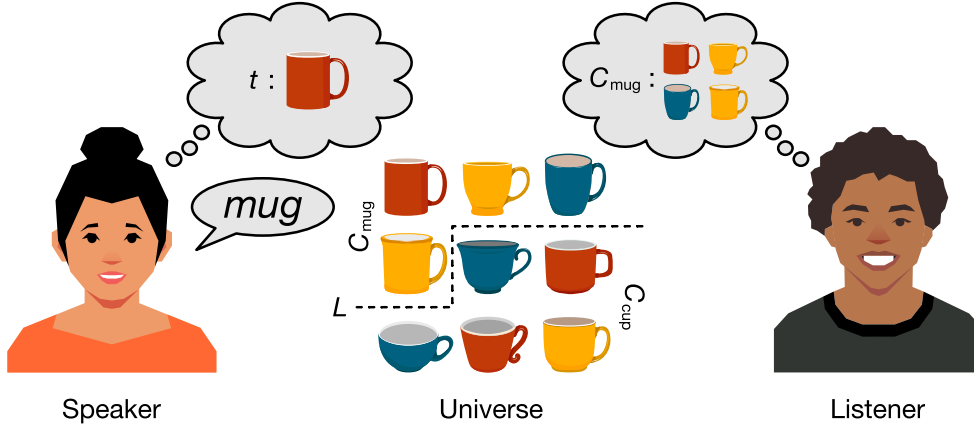
of other domains, including numeral systems (Y. Xu & Regier, 2014) and spatial relationship terms (Khetarpal, Neveu, Majid, Michael, & Regier, 2013), studies that have yielded the same basic result: In large, cross-linguistic datasets, natural languages are both optimally simple and optimally informative. This is, of course, strikingly similar to the position articulated by Kirby et al. (2015) that we saw back in Chapter 1, although Kirby et al. (2015) have, in addition, put forward two pressures that gave rise to this state of affairs (learning and communication), while Regier and colleagues have tended to steer clear of positing particular mechanisms. Later, in Section 2.4, I reflect on how these two bodies of literature – coming at the issue from somewhat different directions – relate to each other. First, however, we turn our attention to communicative cost.

### 2.3.2 Communicative cost

This section describes Regier and colleagues' information-theoretic measure of informativeness, communicative cost, which plays a crucial role in Chapter 3. Here I provide a fairly detailed description of the measure from a coding perspective, expanding on Regier and colleagues' standard description with the goal of making Chapter 3 more accessible. We will also see how the definition of communicative cost yields various predictions about the features we would expect to find in informative languages.

The central idea behind communicative cost is that languages may be described as informative to the extent that, during communicative interaction, they minimize information loss. Communicative interaction inherently involves a loss of information because, while the speaker may be certain about the meaning to be expressed, the listener only has access to a word that describes a general category of meanings. There are two main formalizations of communicative cost, which are adopted according to the particular semantic domain being modelled. I begin by describing the simple discrete case and then move on to the continuous case.

**Discrete categories** In Kemp and Regier's (2012) study of kinship terms (described above), the discrete form of communicative cost is adopted. Here I give an example where a simple language divides nine possible drinking vessels into two categories, cups and mugs. A communicative interaction is illustrated in Fig. 2.9. There is a universe of meanings  $U$ , and the speaker and listener have a shared language  $L$  that partitions



**Figure 2.9:** A speaker wishes to communicate a target meaning  $t$  from a universe of meanings  $U$ . She determines which category the target belongs to according to the shared language  $L$ , which partitions  $U$  into categories, and utters its associated word. The listener maps this word back to a category of possible meanings and must decide on a specific meaning to infer. Here the listener has a  $1/4$  chance of correctly inferring the target, since there are four members of the mug category.

$U$  into categories:  $L = \{C_{\text{cup}}, C_{\text{mug}}\}$ . The speaker wishes to communicate a target meaning  $t \in U$ , so she determines which category the target belongs to, for example  $C_{\text{mug}}$ , and transmits its associated word to the listener. The listener maps the word back to the category and selects a meaning  $t' \in C_{\text{mug}}$ . The interaction is considered successful if  $t = t'$ . In general, the probability that the speaker will successfully communicate some meaning  $m$  using signal  $s$  as an intermediary is  $1/|C_{s(m)}|$ ; as the cardinality of the category grows, the probability of success decreases because the listener becomes less certain about the speaker's intended meaning. The loss of information, or cost, incurred by using a signal as a proxy for a meaning is therefore  $-\log 1/|C_{s(m)}|$  bits; put differently, the cost is how much additional information the speaker must still send (e.g. additional modifiers) for the listener to pick out the intended meaning.

The cost of sending a meaning is modulated by the probability of that meaning occurring  $p(m)$ , which Regier and colleagues refer to as the 'need probability'. Thus, the expected cost of a language as a whole is given by

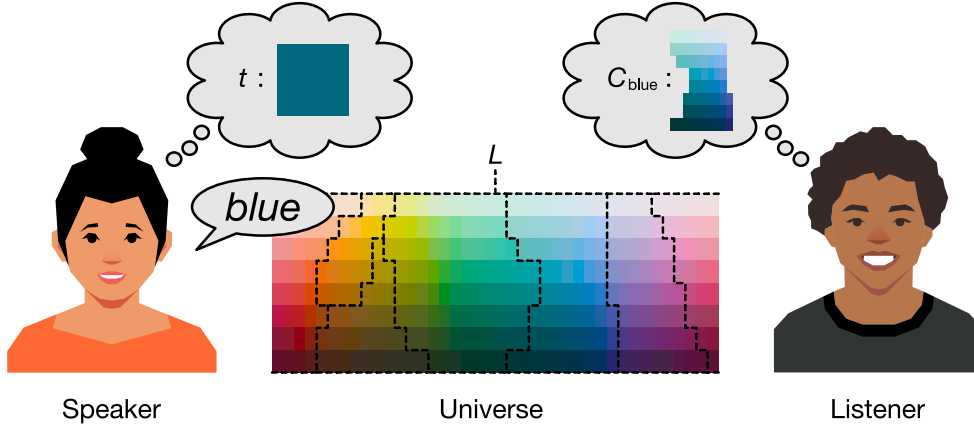
$$\text{cost}(L) = \sum_{m \in U} p(m) \cdot -\log \frac{1}{|C_{\text{signal}(m)}|}. \quad (2.8)$$

For each meaning in the universe, we multiply the probability of that meaning occurring by how much information is lost when a category/signal is used as a proxy for that

**Table 2.1:** Optimal lossless codewords and lossy signals used to represent 64 meanings

Meaning	Lossless codeword	Category	Lossy signal	Meaning	Lossless codeword	Category	Lossy signal
$m_1$	000000	$C_1$	00	$m_{33}$	100000	$C_3$	10
$m_2$	000001	$C_1$	00	$m_{34}$	100001	$C_3$	10
$m_3$	000010	$C_1$	00	$m_{35}$	100010	$C_3$	10
$m_4$	000011	$C_1$	00	$m_{36}$	100011	$C_3$	10
$m_5$	000100	$C_1$	00	$m_{37}$	100100	$C_3$	10
$m_6$	000101	$C_1$	00	$m_{38}$	100101	$C_3$	10
$m_7$	000110	$C_1$	00	$m_{39}$	100110	$C_3$	10
$m_8$	000111	$C_1$	00	$m_{40}$	100111	$C_3$	10
$m_9$	001000	$C_1$	00	$m_{41}$	101000	$C_3$	10
$m_{10}$	001001	$C_1$	00	$m_{42}$	101001	$C_3$	10
$m_{11}$	001010	$C_1$	00	$m_{43}$	101010	$C_3$	10
$m_{12}$	001011	$C_1$	00	$m_{44}$	101011	$C_3$	10
$m_{13}$	001100	$C_1$	00	$m_{45}$	101100	$C_3$	10
$m_{14}$	001101	$C_1$	00	$m_{46}$	101101	$C_3$	10
$m_{15}$	001110	$C_1$	00	$m_{47}$	101110	$C_3$	10
$m_{16}$	001111	$C_1$	00	$m_{48}$	101111	$C_3$	10
$m_{17}$	010000	$C_2$	01	$m_{49}$	110000	$C_4$	11
$m_{18}$	010001	$C_2$	01	$m_{50}$	110001	$C_4$	11
$m_{19}$	010010	$C_2$	01	$m_{51}$	110010	$C_4$	11
$m_{20}$	010011	$C_2$	01	$m_{52}$	110011	$C_4$	11
$m_{21}$	010100	$C_2$	01	$m_{53}$	110100	$C_4$	11
$m_{22}$	010101	$C_2$	01	$m_{54}$	110101	$C_4$	11
$m_{23}$	010110	$C_2$	01	$m_{55}$	110110	$C_4$	11
$m_{24}$	010111	$C_2$	01	$m_{56}$	110111	$C_4$	11
$m_{25}$	011000	$C_2$	01	$m_{57}$	111000	$C_4$	11
$m_{26}$	011001	$C_2$	01	$m_{58}$	111001	$C_4$	11
$m_{27}$	011010	$C_2$	01	$m_{59}$	111010	$C_4$	11
$m_{28}$	011011	$C_2$	01	$m_{60}$	111011	$C_4$	11
$m_{29}$	011100	$C_2$	01	$m_{61}$	111100	$C_4$	11
$m_{30}$	011101	$C_2$	01	$m_{62}$	111101	$C_4$	11
$m_{31}$	011110	$C_2$	01	$m_{63}$	111110	$C_4$	11
$m_{32}$	011111	$C_2$	01	$m_{64}$	111111	$C_4$	11

meaning. This definition of communicative cost has a natural interpretation in information theory: It is the expected number of additional bits required to unambiguously encode a meaning beyond the number of bits that were actually transmitted. An example is illustrated in Table 2.1: A universe consists of 64 equally probable meanings, such that the lossless codeword for each meaning would optimally require  $\log 64 = 6$  bits, but the language divides the meanings into four equally-sized categories, such that the signal used to represent each category requires just  $\log 4 = 2$  bits (00, 01, 10, or 11). Thus, in this example, the communicative cost would be  $6 - 2 = 4$  bits, since on every attempt to communicate, four bits of information is lost. Another way to think about this is that the lossy signal uttered by the speaker only consists of the first two digits of the ideal lossless codeword. In essence, then, communicative cost quantifies the average amount of information that will be lost under a lossy – as opposed to lossless –

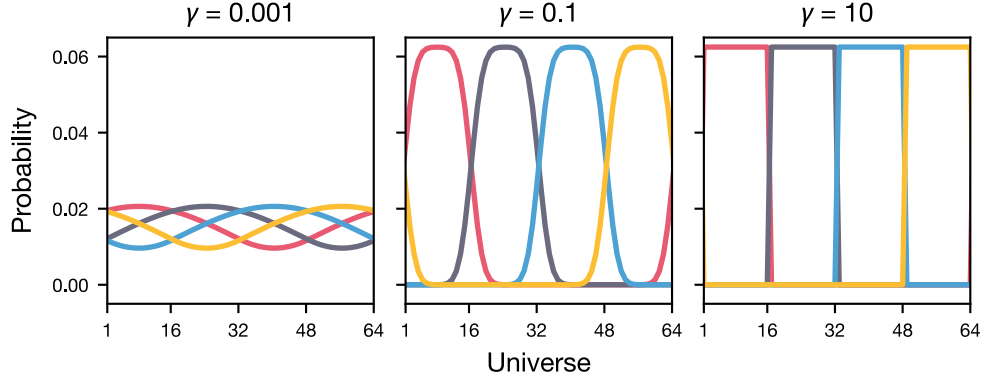


**Figure 2.10:** A speaker wishes to communicate a target meaning  $t$  from a universe of meanings. She determines which category the target belongs to according to the shared language  $L$ , which partitions the universe into categories, and utters its associated word. The listener maps this word back to a category of possible meanings and must decide on a specific meaning to infer. He is more likely to infer prototypical (central) members of the category.

coding system, and is therefore a kind of negative Kullback–Leibler divergence.<sup>8</sup>

**Continuous categories** The definition of communicative cost above assumes that, on hearing a word, the listener is equally likely to infer each member of the associated category. As such, it is only sensitive to how many categories are available and the extent to which category sizes reflect the need probabilities. While this may be appropriate in cases where categories are discrete – Kemp and Regier’s (2012) study of kinship terms, for example – it is overly-simplistic in others because it fails to model two well-known effects of human categorization: Categories may have fuzzy boundaries, and some category members may be more prototypical than others. Therefore, in work with continuous meanings (such as the colour term study in Regier et al., 2015), the following extended notion of communicative cost is used, in which the probability of successfully transmitting a meaning from one mind to another is not simply  $1/|C|$  but a value related to the distance between intended and inferred meanings. An example of this is illustrated in Fig. 2.10.

<sup>8</sup> The Kullback–Leibler divergence,  $D_{KL}(P||Q) = \sum_i P(i) \log P(i)/Q(i)$ , quantifies the expected information loss when a lossless code optimized for probability distribution  $Q$  is used to encode samples from  $P$  rather than using a lossless code optimized for  $P$ . This coding interpretation of the Kullback–Leibler divergence is similar to communicative cost, except communicative cost compares a lossless code against a lossy code for the same distribution, resulting in a negative divergence.



**Figure 2.11:** A one-dimensional space of 64 meanings is broken up into four equally sized categories. Each probability distribution (indicated in different colours) represents the probability that a listener would infer a given meaning on hearing the word for a given category. The listener is more likely to infer meanings at the centre of the category because they are more prototypical. Each plot shows these probability distributions under different settings of  $\gamma$ . When  $\gamma$  is small, category boundaries are very fuzzy; when  $\gamma$  is large, the distributions collapse to the discrete case.

Rather than treat each category  $C$  as a *set* of meanings, each category will now be treated as a probability distribution over all  $m \in U$ . To transform a set  $C$  into a distribution  $\tilde{C}$ , we assume that the probability of inferring a meaning is proportional to its total similarity to all category members  $m' \in C$ :

$$\tilde{C}(m) \propto \sum_{m' \in C} \exp -\gamma d(m, m')^2, \quad (2.9)$$

where  $\gamma > 0$  and  $d(\cdot, \cdot)$  gives the distance between meaning  $m$  and meaning  $m'$ . The term  $\exp -\gamma d(m, m')^2$  relates distance to perceived similarity: The similarity between a meaning and itself is 1; as the distance between two meanings grows, the similarity approaches 0. The parameter  $\gamma$  controls how quickly similarity decays with distance. The effect of this parameter is illustrated in Fig. 2.11. Essentially, small values model fuzzier category systems in which the boundaries between categories are blurred; as  $\gamma$  becomes arbitrarily large,  $\tilde{C}(m) = 1/|C|$  if  $m \in C$  and 0 otherwise, collapsing to the discrete case described above.

The transformation performed by Equation 2.9 models the categories as Gaussians in which the most prototypical meaning (the meaning at the geometric centre of the category with the greatest similarity to other category members) has the highest probability of being inferred. Since  $\tilde{C}(m)$  gives the probability that the listener will infer meaning  $m$



on hearing the signal for category  $\tilde{C}$ , the cost of sending that meaning is  $-\log \tilde{C}(m)$ . Therefore, the communicative cost of the language as a whole is given by

$$\text{cost}(L) := \sum_{m \in U} p(m) \cdot -\log \tilde{C}_{\text{signal}(m)}(m). \quad (2.10)$$

### 2.3.3 Hallmarks of informative category systems

According to communicative cost, the extent to which a language or category system may be described as informative is determined by four basic properties. In the discrete case, the following two properties hold:

1. **Expressivity:** The more categories a system has, the lower its communicative cost; a system of many categories is informative because the categories pick out smaller, more precise sets of meanings.
2. **Balanced cardinality:** The more evenly-balanced the category sizes are, the lower the communicative cost of the system as a whole.

In the continuous case, the following property also applies:

3. **Compactness:** The more compact categories are, the lower the communicative cost of the system; compact categories are informative because they minimize the distance between intended and inferred meanings.

This property arises directly out of the assumptions encoded into Equation 2.9. By assuming that categories are nonuniform – that some meanings are better examples of a category than others – communicative cost builds in a preference for compactness; it pays for a category to be compact because if a listener is more likely to guess more prototypical meanings in response to hearing the word for a given category, then arranging the category such that category members are generally close to the prototypical meaning maximizes the chance that speaker and listener will infer the same meaning. This is a convoluted way of saying that an informative category system is one in which the speaker's intended meaning and the listener's inferred meaning are, on average, as similar as possible. However, communicative cost is not formulated directly in this way; rather, Regier and colleagues typically describe it in terms of the loss of information

between two distributions or representations – the speaker’s representation of the target meaning and the listener’s representation of what that meaning might be based on the word they have received. But since the listener’s representation of a category is estimated from similarity-based assumptions about prototypicality, communicative cost builds in the compactness property.<sup>9</sup>

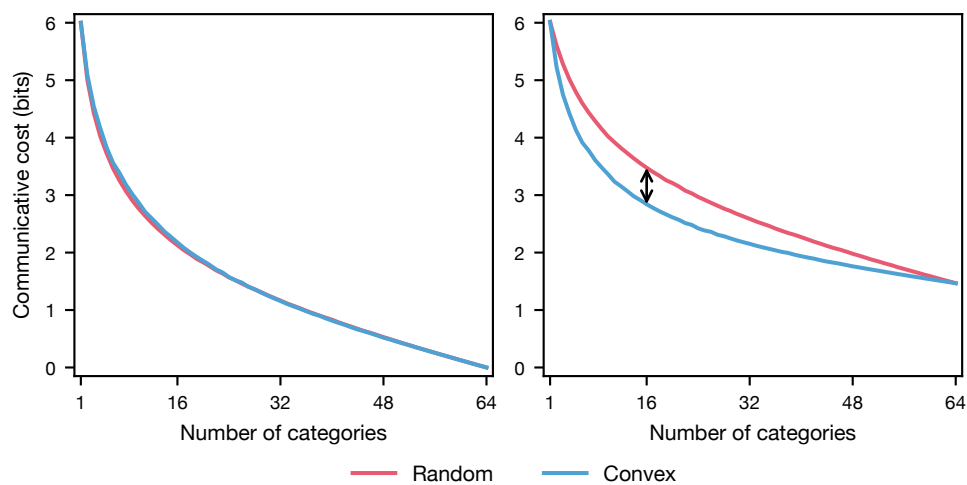
Finally, if the need probabilities are nonuniform, the properties above may be overridden in favour of smaller categories for frequent meanings and larger categories for infrequent meanings (see e.g. Gibson et al., 2017; Regier, Carstensen, & Kemp, 2016). This leads to a fourth property of informativeness:

4. **Reflection of need:** When categories reflect the needs of the interlocutors, the communicative cost of the system is lowered.

Of course, these four hallmark features of informativeness are correlated in various ways. For example, in a system that has more categories, those categories will automatically be more compact, and in a system where compactness is maximized, the cardinality of categories will tend to be balanced. However, the important point I want to emphasize here is that communicative cost – a measure of informativeness – is determined by various underlying factors.

For the purpose of this thesis, I will focus only on the expressivity and compactness properties. To demonstrate that these properties are considered informative under communicative cost, we can reconsider the random and convex category systems that were simulated in Section 2.2.6 (page 32; see also Fig. 2.2 on page 19). Fig. 2.12 plots the communicative cost of these systems as a function of  $n$  under both the discrete measure (left) and the continuous measure (right). In either case, the more categories a system has, the more informative it is (the lower its communicative cost); greater expressivity makes a language more informative. Secondly, for a given number of categories, convex systems (blue) are more informative than random systems (pink), but only under the continuous form of communicative cost, as highlighted by the black arrow in Fig. 2.12. In other words, the same compactness advantage that we observed in terms of simplicity (see page 33) also exists in terms of informativeness.

<sup>9</sup> It is unclear whether this was intended by Regier and colleagues; their work typically describes Equation 2.9 as providing a Gaussian model of a category (e.g. Regier et al., 2015, p. 244) and does not appear to recognize that this also builds in a compactness advantage. Nevertheless, I would contend that a good measure of informativeness should indeed value compactness in this way.



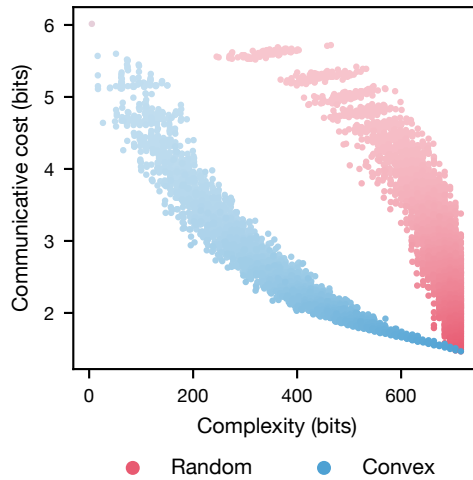
**Figure 2.12:** Informativeness of simulated category systems in an 8×8 space under the discrete measure of communicative cost (left) and the continuous measure (right). The pink curves show the mean communicative cost of randomly generated category systems as a function of the number of categories. The more categories a system has, the less costly it is to use. The blue curves show equivalent results for *convex* category systems. Systems that consist of compact categories are less costly than unstructured systems, but only when the continuous measure (right) is adopted; under the discrete measure, there is no such compactness advantage.

## 2.4 The Simplicity–Informativeness Tradeoff

The task of category systems is to provide maximum information with the least cognitive effort.

— Eleanor Rosch (1978)

The body of research reviewed above hints at two converging literatures. Regier and colleagues focus on the typological distribution of languages – showing that real languages are optimally simple and informative – but they abstract away from the mechanistic details (Levinson, 2012, p. 989). Meanwhile, Kirby and colleagues attempt to be more explicit about how cognitive principles and cultural dynamics are linked to the design features of language; in other words, they attempt to solve what Kirby (1999) called ‘the problem of linkage’. These two bodies of work are complementary and – as we have seen – have arrived at similar conclusions, conclusions that are succinctly summarized by the quotation from Rosch above. The job of a category system is to permit informative communication in the simplest possible way, and this gives rise to what I call the *simplicity–informativeness tradeoff*:



**Figure 2.13:** Complexity and communicative cost of random (pink) and convex (blue) category systems that have between 1 and 64 categories; the darker the colour, the more categories the system has. This plot combines the results shown in Fig. 2.7 and Fig. 2.12. Together, the points delimit the space of possible languages in simplicity–informativeness space. The tradeoff between simplicity and informativeness results in a downward sloping relationship between complexity and cost; complex languages are informative and simple languages are uninformative. Convexity (a form of compactness) is both simple and informative, so convex systems lie along the optimal frontier.

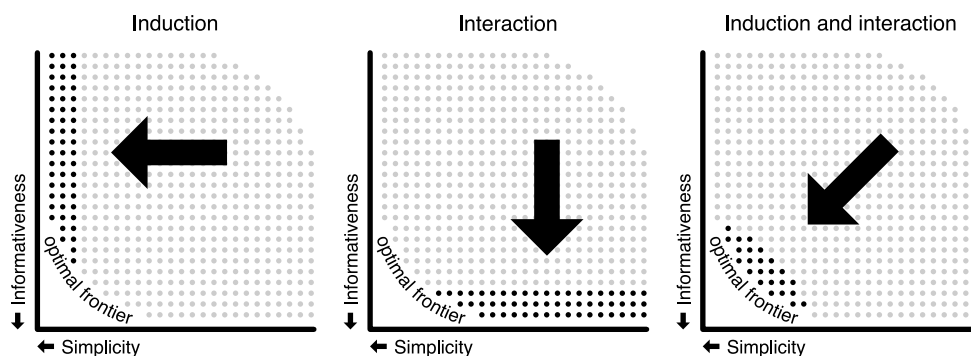
**The simplicity–informativeness tradeoff:** Successful languages are both simple and informative, but, with some caveats, simple languages cannot be informative, and informative languages cannot be simple, giving rise to a tradeoff.

The main caveat is, of course, that the compactness advantage applies to both simplicity (see page 33) and informativeness (see page 43), and since compact structure is both simple and informative, it is not subject to the simplicity–informativeness tradeoff, a point that becomes important in the next chapter.

Combining the simulation results from the previous two sections (refer back to Figs. 2.7 and 2.12) gives us a concrete picture of the tradeoff, as shown in Fig. 2.13, which depicts the results within Kemp and Regier’s (2012, p. 1052) simplicity–informativeness space.<sup>10</sup> The downward sloping relationship between complexity and cost results from the fact that simple category systems are generally not informative and informative category systems are generally not simple. However, note that the convex systems in blue, which have a compact structure and are therefore favoured by both learning and communication, lie along the optimal frontier.

Returning to Kirby et al. (2015), who posit that learning acts as the pressure for simplicity and communicative interaction acts as the pressure for informativeness, yields a view of the evolution of conceptual systems that is illustrated in Fig. 2.14: When there is only a pressure from induction, languages become simple, moving westwards in simplicity–informativeness space; when there is only a pressure from interaction, lan-

<sup>10</sup> See also Kemp, Xu, and Regier (2018) who similarly describe movements in this space.



**Figure 2.14:** Illustration of how pressures from induction and interaction are expected to act in simplicity–informativeness space. When there is only a pressure from induction, languages become simple; when there is only a pressure from interaction, languages become informative; when both pressures are in play, the languages adapt to become both simple and informative, clustering around the optimal frontier.

languages become informative, moving south in simplicity–informativeness space; when both pressures are in play, the languages adapt to become both simple and informative, clustering around the optimal frontier. Systems at the optimal frontier ought to have a compact structure, since compactness is favoured by both pressures, but they will find some tradeoff in terms of expressivity; too few categories, and the language is not informative enough, but too many categories and the language is not simple enough.

## 2.5 Conclusion to Chapter 2

In this chapter I have introduced three broad areas of research that are relevant to this thesis. First, in Section 2.1, we reviewed a few key ideas in the concepts and categorization literature. In Section 2.2, we formalized a model of learning in terms of Bayesian induction under a prior bias for simplicity. By applying a simplicity principle to the problem of inducing a grammar from incomplete, noisy data, the rational learner is able to maximize the probability of correctly inferring how the world truly works. This formalization is put to the test in the next chapter. In Section 2.3, we reviewed a recent body of work from Regier and colleagues, which has convincingly demonstrated in a variety of domains that natural languages appear to be well optimized in terms of both simplicity and informativeness. Furthermore, this body of literature provides an information-theoretic formalization of informativeness, called communicative cost.

	Pressure from learning/induction	Pressure from communication/interaction
Kirby et al.	<b>compressible languages</b> measured by: <i>complexity</i>	<b>expressive languages</b> measured by: <i>number of words</i>
Regier et al.	<b>simple category systems</b> measured by: <i>number of words</i>	<b>informative category systems</b> measured by: <i>communicative cost</i>
this thesis	<b>simple languages</b> measured by: <i>complexity</i>	<b>informative languages</b> measured by: <i>communicative cost</i>

**Figure 2.15:** The basic paradigms promoted by Kirby and colleagues and Regier and colleagues. Kirby and colleagues talk about a tradeoff between compressible and expressive languages deriving from learning and communication; there is a special emphasis on (iterated) learning which is formalized in terms of information-theoretic complexity, while expressivity is approximated by the number of words. Regier and colleagues talk about a tradeoff between simple and informative communication systems, which maps fairly neatly onto the view from Kirby and colleagues. There is a special emphasis on informativeness, which is formalized as communicative cost, while simplicity is approximated as the number of words. This thesis attempts to unify these two paradigms.

This formalization is especially useful because it fills a gap that has existed thus far in the iterated learning literature – a more formal account of expressivity. In this thesis I adopt a combination of the two paradigms, as highlighted in Fig. 2.15. Finally, in Section 2.4, we saw how the tradeoff between learning and interaction are expected to play out in terms of conceptual structure and that compactness has special status under the tradeoff because it is a feature of both simplicity and informativeness.

## Chapter 3

# Simplicity from Induction

We are to admit no more causes of natural things, than such as are both true and sufficient to explain their appearances. To this purpose the philosophers say, that Nature do's nothing in vain, and more is in vain, when less will serve; For Nature is pleas'd with simplicity, and affects not the pomp of superfluous causes.

— Isaac Newton (1729)

Newton's first rule of reasoning is an expression of Occam's razor: All things being equal, simpler explanations should be preferred over more complex ones because 'Nature do's nothing in vain'. This is an especially apt place to begin this chapter because we will apply Occam's razor on two levels. Firstly, we will instantiate Bayesian agents with an Occam's razor prior – confronted with noisy, incomplete data these agents look for simple hypotheses to explain that data. But secondly, the argument itself also appeals to Occam's razor by contending that a simplicity preference offers a more parsimonious explanation of compact conceptual structure than a reasonable alternative that has been put forward in the literature.

In a commentary on Kemp and Regier (2012), the study of kinship systems that we looked at the previous chapter, Levinson (2012, p. 989) pointed out that, although their findings demonstrate that real languages are optimized in terms of both simplicity and informativeness, the work does not provide an explanation for 'where our categories come from'. In other words, the typological studies on the simplicity–informativeness tradeoff that we reviewed in Section 2.3.1 describe how languages *are*, but they do not offer explanations for *why* languages should be that way. Levinson (2012, p. 989) went

on to suggest a few ways in which this issue could be tackled, including through experimental approaches that ‘show how categories get honed through iterated learning across simulated generations’.

In direct response to this comment, Carstensen, Xu, Smith, and Regier (2015) conducted two studies of the iterated learning of category structures, and these studies have – in our view – yielded a surprising result: The authors found that *informative* category systems can arise through iterated learning without any communicative pressure. I describe this as surprising for two reasons. Firstly, the result runs contrary to the literature reviewed in the previous two chapters, which argue that learning – and therefore iterated learning – have a simplifying effect on language. Intuitively, and as demonstrated in Section 2.4, one would expect to find that informative languages have greater complexity, so at first glance it is unclear why iterated learning would yield informative category structures. Secondly, the result also appears to contradict Regier and colleagues’ own approach to informativeness, which is very much construed in communicative terms. Communicative cost, for example, is a measure of how much information is lost during communicative interaction, so, at first glance, it is unclear why the authors would take the position that an increase in informativeness (a decrease in communicative cost) would occur through pressure from learning.

Carstensen et al.’s (2015) position appears to be that learners expect languages to be informative and are therefore equipped with a bias for informative languages; the effects of this bias are then amplified by the process of iterated learning. This is certainly not an unreasonable claim, and similar ideas can be found in Fedzechkina, Jaeger, and Newport (2012) and Frank and Goodman (2014). Paper 2, which is the main content of this chapter, is a direct response to Carstensen et al. (2015). In the paper, we formalize two positions about the content of the human inductive bias in terms of two possible priors: a prior for simplicity or a prior for informativeness. We then test the predictions of these formalizations experimentally.

### 3.1 Preface to Paper 2

Paper 2 was under review at the time of the submission of this thesis; the manuscript reproduced over the following pages was submitted to the *PsyArXiv* preprint server on 1 July 2018 (<https://doi.org/10.31234/osf.io/jkfyx>; Version 1). The citations may be



looked up on pages 83–86 or in the references list at the end of this volume. The paper makes reference to four supplementary items, which may be located as follows:

- S1. *An introduction to communicative cost*: Section 2.3.2, page 38.
- S2. *All model results*: Appendix A, page 159.
- S3. *Participant exclusion and attrition*: Appendix B, page 177.
- S4. *Individual participant results in Experiment 1*: Appendix C, page 181.

All work reported in the paper is my own, including the technical development of the model and experiments. The contributions made by my coauthors were as follows:

**Kenny Smith** Advice on model design, experimental design, statistical methods, the model fit procedure, and general editing of the paper.

**Jennifer Culbertson** Advice on experimental design and statistical methods, and general editing of the paper.

**Simon Kirby** Conception of the basic idea behind the paper, advice on model design and experimental design, and general editing of the paper.

## Simplicity and informativeness in semantic category systems

Jon W. Carr, Kenny Smith, Jennifer Culbertson, Simon Kirby

School of Philosophy, Psychology and Language Sciences, University of Edinburgh

### Abstract

Recent research has shown that semantic category systems, such as color and kinship terms, find an optimal balance between the considerations of simplicity and informativeness. We argue that this situation arises through a pressure for simplicity from learning and a pressure for informativeness from communicative interaction, two distinct pressures that pull in (often but not always) opposite directions. An alternative account suggests that learning might also act as a pressure for informativeness—that learners might be biased toward inferring informative systems. This results in two competing hypotheses about the human inductive bias. We formalize these competing hypotheses in a Bayesian iterated learning model and test them in two experiments with human participants. Specifically, we investigate whether learners’ inductive biases, isolated from any communicative task, are better characterized as favoring simplicity or informativeness. We find strong evidence to support the simplicity account. Furthermore, we show how the application of a simplicity principle in learning can give the impression of a bias for informativeness, even when no such bias is present. Our findings suggest that semantic categories are learned through domain-general principles, negating the need to posit a domain-specific inductive bias.

*Keywords:* category learning; induction; informativeness; iterated learning; language evolution; simplicity

### Introduction

We make sense of the world through a rich system of learned concepts, which allow us to categorize and make predictions about an infinite range of perceptual stimuli on the basis of their similarities to previous encounters (Gärdenfors, 2014; Lakoff, 1987; Murphy, 2004; Rosch, 1973; Shepard, 1987). However, there is no singular, objective way of conceptualizing the world, and different human populations align on different systems of categorization. In the domain of kinship, for example, different languages have different ways of grouping

---

Correspondence should be sent to Jon Carr, School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Dugald Stewart Building, 3 Charles Street, Edinburgh, EH8 9AD, United Kingdom. Email: j.w.carr@ed.ac.uk

## SIMPLICITY AND INFORMATIVENESS IN SEMANTIC CATEGORY SYSTEMS 2

family members into labeled categories (Murdock, 1970). Cantonese, for example, makes a lexical distinction between all four grandparents—yèh (father’s father), màh (father’s mother), gūng (mother’s father), and pòh (mother’s mother) (Cheung, 1990), while most varieties of English collapse the distinction between the maternal and paternal lineage.

Despite the diversity in how languages classify meanings into categories, there is evidence to suggest that these systems achieve a balance between simplicity (e.g., the number of categories that must be learned) and informativeness (e.g., the ability to express needed distinctions). Different languages find different solutions to this tradeoff (as in the comparison between Cantonese and English above), but they nevertheless offer a near optimal balance between the two considerations. This has been shown in a variety of domains, including kinship terms (Kemp & Regier, 2012), spatial relationships (Khetarpal, Neveu, Majid, Michael, & Regier, 2013), numeral systems (Y. Xu & Regier, 2014), color terms (Regier, Kemp, & Kay, 2015), and container names (Y. Xu, Regier, & Malt, 2016).

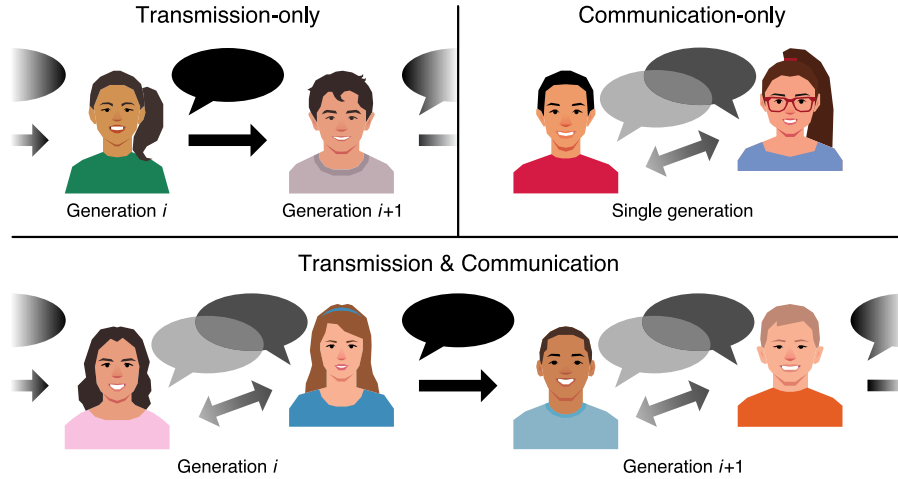
The idea that languages are shaped by opposing forces has a long history in linguistics and cognitive science; such ideas can be found in the works of von der Gabelentz (1891), Zipf (1949), and Martinet (1952). The notion of a tradeoff between simplicity and informativeness in semantic categories is often attributed to Rosch (1978, p. 28) who argued that “the task of category systems is to provide maximum information with the least cognitive effort,” a view echoed by Gärdenfors (2014, p. 132–133) who suggests that concepts achieve “a balance between the precision of the noun and the number of words that have to be remembered.” Kemp, Xu, and Regier (2018, p. 111) make this even more explicit:

These two desiderata [simplicity and informativeness] necessarily compete against each other. A highly informative communicative system would be very fine-grained, detailed, and explicit—and would as a result be complex, not simple. A very simple system, in contrast, would necessarily leave implicit or unspecified many aspects of the speaker’s intended meaning—and would therefore not be very informative. A system supports efficient communication to the extent that it achieves an optimal trade-off between these two competing considerations.

However, as Levinson (2012, p. 989) has pointed out, although this body of research demonstrates that natural languages are both simple and informative and that a tradeoff exists between these properties, it does not specify the mechanisms that give rise to this state of affairs. Two recent strands of research attempt to resolve this.

The first strand posits two distinct pressures that play out during language learning and language use. The pressure for simplicity derives from an inductive bias, which leads learners to prefer simpler languages. This is because, when learners are trying to understand the world, the best strategy—given that they have no expectations about how the world is structured—is to apply Occam’s razor; all things being equal, simpler hypotheses should be preferred over more complex hypotheses (Li & Vitányi, 2008; Rissanen, 1978; Solomonoff, 1964). This is a highly general principle, claimed to operate across cognitive domains (Chater, Clark, Goldsmith, & Perfors, 2015; Chater & Vitányi, 2003; Culbertson & Kirby, 2016; Feldman, 2016; Kemp, 2012), but its effect on linguistic and conceptual systems is amplified by *iterated learning*, the process by which learned systems are passed down through generations via cultural transmission. Canini, Griffiths, Vanpaemel, and Kalish

## SIMPLICITY AND INFORMATIVENESS IN SEMANTIC CATEGORY SYSTEMS 3



*Figure 1.* Three models of the cultural evolution of language. In Transmission-only, a language is repeatedly learned and transmitted to a new generation, exerting a pressure for simplicity which accumulates over generations. In Communication-only, a pair of interlocutors repeatedly interact with each other, which exerts a pressure for informativeness. Transmission & Communication combines pressures for both simplicity and informativeness.

(2014), for example, show that iterated learning replicates a number of findings from classic research in concept learning, such as the bias for marking distinctions on one dimension rather than two (e.g., Ashby & Maddox, 1990; Moreton, Pater, & Pertsova, 2015; Shepard, Hovland, & Jenkins, 1961). However, on its own, an inductive bias for simplicity will ultimately result in languages that are maximally simple; in the limit, iterated learning converges to the prior (Griffiths & Kalish, 2007).

According to this first strand of research, the process that keeps overly simple languages at bay is communicative interaction. Kirby, Tamariz, Cornish, and Smith (2015) directly investigate the distinct roles of learning and interaction with three models of the cultural evolution of language (see Fig. 1). In the Transmission-only model, a language is transmitted from one individual to the next in a chain of learners. As expected, repeated learning of the language gives rise to simple systems (e.g., systems that divide meanings into relatively few, more general categories). In the Communication-only model, two individuals repeatedly interact with each other with the goal of successfully communicating. These repeated interactions give rise to highly complex languages in which every meaning has its own unique word, allowing the interlocutors to achieve maximum success. The Transmission & Communication model combines the two pressures: At each generation a pair of interlocutors attempt to communicate (imposing the pressure for informativeness), but the language they produce is then learned by a new pair of interlocutors (imposing the pressure for simplicity). In this case, the resultant languages develop compositional structure that is both easy to learn (simple) and fully productive (informative); working in tandem,

## SIMPLICITY AND INFORMATIVENESS IN SEMANTIC CATEGORY SYSTEMS 4

the two pressures result in languages that find an optimal balance between simplicity and informativeness. Similar findings regarding the roles of learning and interaction have been observed in the gestural modality (Motamedi, Schouwstra, Culbertson, Smith, & Kirby, in press) and under more complex meaning spaces (Carr, Smith, Cornish, & Kirby, 2017). See Kirby, Griffiths, and Smith (2014) and Tamariz (2017) for reviews.

The second strand of research explains the pressure for informativeness in terms of an inductive principle rather than the dynamics present in interaction. Focusing on a particular formalization of informativeness, called *communicative cost*,<sup>1</sup> Carstensen, Xu, Smith, and Regier (2015) provide evidence that informative semantic categories can arise through iterated learning, suggesting that humans may have an inductive bias that favors more informative systems. The authors conducted two studies: Study 1 reanalyzed data from an iterated learning experiment with color terms (J. Xu, Dowman, & Griffiths, 2013), and Study 2 was a novel iterated learning experiment using spatial relationship stimuli. Both studies demonstrated an increase in informativeness over generations in a transmission-only design, crucially involving no interaction. These results are at odds with those described above which document the emergence of degenerate or uninformative languages in the absence of a shared communicative task (Carr et al., 2017; Kirby et al., 2015) or an artificial analog of such a task (Beckner, Pierrehumbert, & Hay, 2017; Kirby, Cornish, & Smith, 2008).

However, the idea that learning in the absence of communicative interaction can lead to informative systems has also been suggested by Fedzechkina, Jaeger, and Newport (2012). In their studies, adult participants restructure miniature artificial languages in ways that appear to balance effort and ambiguity avoidance (i.e., informativeness). In their Experiment 1, for example, participants were trained on a language with variable word order (SOV or OSV) and optional case marking on objects (which were either animate or inanimate). Crucially, the languages that participants were taught were designed to be suboptimal in terms of informativeness—when both event participants were animate, and no case marking was present, it is ambiguous which is the subject and which the object. A more informative system (which is not overly complex) would consistently use case marking with animate objects, in order to avoid this potential ambiguity. They found that learners inferred just such languages, increasing the case marking on animates and decreasing it on inanimates. Fedzechkina et al. (2012, p. 17900) conclude that, “...language learners are biased toward communicatively efficient linguistic systems and restructure the input language in a way that facilitates information transfer.”

To summarize, we have sketched two theories of the role of learning in the emergence of linguistic and conceptual systems. In the first, learning results in a pressure for simpler systems, while the pressure for informativeness comes from communicative interaction. In the second, informativeness is directly built into the process of learning. Both theories have experimental evidence to support them, but have not been directly compared. In this paper we describe a model of a Bayesian category learner which implements these alternative hypotheses in terms of two possible inductive prior biases: a bias for simplicity or a bias for informativeness. We show what kinds of languages are expected to result from iterated learning under each. We then test the predictions of the model in two experiments. In an objective model comparison, we find that human learning biases are better

<sup>1</sup>This is explained in more detail later, but see Regier et al. (2015), Kemp et al. (2018), or supplementary item S1 for a more complete introduction.

## SIMPLICITY AND INFORMATIVENESS IN SEMANTIC CATEGORY SYSTEMS 5

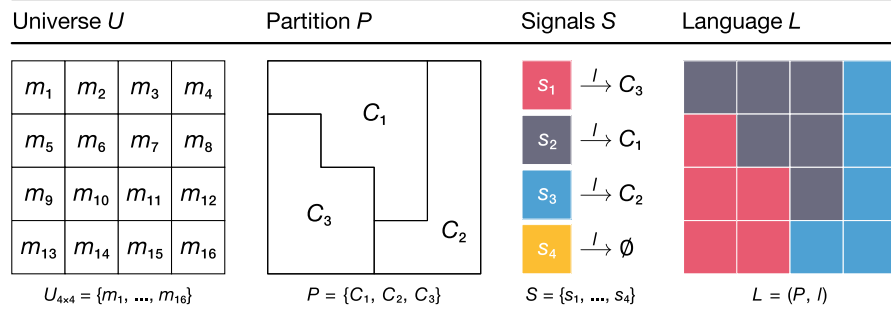


Figure 2. A universe is a two-dimensional metric space consisting of  $M$  meanings. This space may be partitioned into  $N$  mutually disjoint categories. Agents are provided with a fixed set of  $N_{\max}$  signals  $S$ , which are used to label the categories. In this example, the language partitions a  $4 \times 4$  space into three contiguous categories using three of the four available signals (signals are indicated by colors).

characterized by a preference for simplicity. Furthermore, we show how the application of a simplicity principle in learning can explain the apparently contradictory experimental results of Carstensen et al. (2015). Specifically, under a simplicity prior, iterated learning gives rise to simple category structures that just so happen to have one of the hallmarks of informativeness—*compactness*—which we outline in more detail shortly. We argue that this offers an alternative, more parsimonious explanation of Carstensen et al.’s (2015) findings.

### Model

In this section we describe a Bayesian iterated learning model that simulates what happens when a language is passed down a chain of learners. This reveals the kinds of languages that arise when pressure from learning—*on its own*—is repeatedly applied. By manipulating the prior function, we can test two extreme positions about how learning shapes conceptual structure. The first position states that learning imparts a pressure for simplicity from the principle of Occam’s razor; the second position states that learners have a bias for inferring informative categories which imparts a pressure for informativeness. All code and data for our model and experiments are available from <https://osf.io/hkxqp>

### Method

The goal of the learner is to infer how a language partitions a universe of meanings into categories and how those categories are labeled. The basic model framework is illustrated in Fig. 2. The universe consists of  $M$  meanings  $U = \{m_1, \dots, m_M\}$ , which we treat as a metric space  $(U, d)$ , where  $d$  is the distance function defined between meanings. Usually we will refer to this space simply as  $U$ , sometimes also denoting the dimensionality (e.g.,  $U_{4 \times 4}$  for a  $4 \times 4$  space of  $M = 16$  meanings). A partition  $P = \{C_1, \dots, C_N\}$  divides  $U$  into  $N$  categories, such that  $1 \leq N \leq N_{\max}$  where  $N_{\max} \leq M$  defines some arbitrary limit on the number of categories. Each category  $C$  is a set of meanings such that all categories are

## SIMPLICITY AND INFORMATIVENESS IN SEMANTIC CATEGORY SYSTEMS 6

nonempty ( $C_i \neq \emptyset$ ), no meaning exists outside a category ( $\bigcup_{i=1}^N C_i = U$ ), and all categories are mutually disjoint ( $C_i \cap C_j = \emptyset$  for  $i \neq j$ ). Each category is labeled by one signal from a fixed set of  $N_{\max}$  signals  $S = \{s_1, \dots, s_{N_{\max}}\}$  according to a lexicon  $l : S \rightarrow P \cup \emptyset$  (unused signals map to  $\emptyset$  where  $N < N_{\max}$ ). A language is a partition and lexicon,  $L = (P, l)$ , but we use the symbol  $L$  to denote either the partition or the lexicon according to context.<sup>2</sup>

**Likelihood.** The probability of an agent producing signal  $s \in S$  given that it possesses language  $L$  and needs to express meaning  $m$  is given by

$$p(s|L, m; \epsilon) := \begin{cases} 1 - \epsilon & \text{if } m \in L(s) \\ \frac{\epsilon}{N_{\max} - 1} & \text{if } m \notin L(s), \end{cases} \quad (1)$$

where noise on production is controlled by the free parameter  $\epsilon \in (0, 1)$ . If  $\epsilon$  is small, there is a high probability that the agent will produce the correct signal for meaning  $m$  and a low probability that it will produce one of the other  $N_{\max} - 1$  signals at random. During learning, the data observed by an agent is a set of meaning–signal pairs  $D = \{\langle m, s \rangle_1, \langle m, s \rangle_2, \dots\}$ , where meaning  $m$  is labeled by signal  $s$ , a noisy indicator of  $m$ ’s category membership. The likelihood of observing dataset  $D$  if language  $L$  were true is therefore the product of  $p(s|L, m; \epsilon)$  over all meaning–signal pairs:

$$p(D|L; \epsilon) = \prod_{\langle m, s \rangle \in D} p(s|L, m; \epsilon). \quad (2)$$

**Simplicity prior.** The simplicity prior endows agents with an inductive bias favoring simple languages; when the observed data is equally likely under two languages, an agent will prefer the language that is simpler following the principle of Occam’s razor. The simplicity prior  $\pi_{\text{sim}}$  is therefore inversely proportional to the complexity of the language:

$$\pi_{\text{sim}}(L) \propto 2^{-\text{complexity}(L)}. \quad (3)$$

The complexity of a language is given by its description length. For our description method, we adopt Fass and Feldman’s (2002) rectangle code, which provides a set of rectangle “symbols” that may be used to describe an arbitrary region (i.e., a category’s extension) in the universe. Like any alphabet, some symbols are more common than others. Information theory tells us that the optimal codelength of a symbol that occurs with probability  $p$  is  $-\log p$  bits. We follow Fass and Feldman (2002) and assume that rectangle shapes occur with uniform probability, and that, for a given shape, its position in the universe occurs with uniform probability. In  $U_{4 \times 4}$ , this yields the codelengths shown in Table 1 (reproduced from Fass & Feldman, 2002, p. 39) and a total of 100 symbols, which are illustrated in Fig. 3.

A valid description of a category is a set of rectangle symbols that exactly describe the category’s extension, which we call a “rectangularization” of that category. For a given category  $C$ , there are usually many possible rectangularizations, the set of which is given by  $\mathcal{R}(C)$ . A rectangularization  $R \in \mathcal{R}(C)$  that minimizes description length is selected.<sup>3</sup>

<sup>2</sup>The distinction between partition and language is largely unimportant for our purposes. We are mostly concerned with the partition—how the space is structured into discrete categories. Signals merely function as indicators to how the space is partitioned.

<sup>3</sup>In  $U_{4 \times 4}$ , this can be computed quickly, but in larger spaces, the process quickly becomes intractable. We alleviate this using a number of methods. First, a category is separated into independent contiguous chunks,





Table 1

*Calculation of symbol codelengths for a  $4 \times 4$  universe.*

Rectangle shapes	N positions	Probability	Codelength (bits)
$1 \times 1$	16	$1/10 \times 1/16 = 1/160$	$-\log 1/160 = 7.32$
$1 \times 2$	24	$1/10 \times 1/24 = 1/240$	$-\log 1/240 = 7.91$
$1 \times 3$	16	$1/10 \times 1/16 = 1/160$	$-\log 1/160 = 7.32$
$1 \times 4$	8	$1/10 \times 1/8 = 1/80$	$-\log 1/80 = 6.32$
$2 \times 2$	9	$1/10 \times 1/9 = 1/90$	$-\log 1/90 = 6.49$
$2 \times 3$	12	$1/10 \times 1/12 = 1/120$	$-\log 1/120 = 6.91$
$2 \times 4$	6	$1/10 \times 1/6 = 1/60$	$-\log 1/60 = 5.91$
$3 \times 3$	4	$1/10 \times 1/4 = 1/40$	$-\log 1/40 = 5.32$
$3 \times 4$	4	$1/10 \times 1/4 = 1/40$	$-\log 1/40 = 5.32$
$4 \times 4$	1	$1/10 \times 1/1 = 1/10$	$-\log 1/10 = 3.32$

Description lengths are then summed over all categories to obtain the overall complexity of a language:

$$\text{complexity}(L) := \sum_{C \in L} \min_{R \in \mathcal{R}(C)} \sum_{r \in R} -\log p(r). \quad (4)$$

Illustrative examples are shown in Fig. 4.

**Informativeness prior.** To model an inductive bias for informativeness, we directly adopt the communicative cost framework used by Carstensen et al. (2015) and other studies from that literature (Kemp & Regier, 2012; Khetarpal et al., 2013; Regier et al., 2015; Y. Xu & Regier, 2014; Y. Xu et al., 2016). This prior endows agents with an inductive bias toward informative languages; when the observed data is equally likely under two languages, an agent will prefer the language that is more informative. The informativeness prior  $\pi_{\text{inf}}$  is inversely proportional to the communicative cost of the language:

$$\pi_{\text{inf}}(L) \propto 2^{-\text{cost}(L)}, \quad (5)$$

and communicative cost is calculated according to:




$$\text{cost}(L) := \sum_{C \in L} \sum_{m \in C} p(m) \cdot -\log \tilde{C}(m), \quad (6)$$

where  $p(m)$  is the probability of a meaning occurring (assumed to be uniform;  $p(m) = 1/|U|$ ) and  $\tilde{C}(m)$  is the probability that a hypothetical listener would infer meaning  $m$  on hearing the signal associated with category  $C$ . The probability distribution  $\tilde{C}$  is given by

$$\tilde{C}(m) \propto \sum_{m' \in C} \exp -\gamma d(m, m')^2, \quad (7)$$

which are dissected into an initial set of rectangles based on the two chords emanating from each concave vertex. Rectangles that share an edge are then recursively merged until no further mergers are possible using dynamic programming techniques. For large chunks an estimate must be obtained by beam search due to an explosion in the number of candidate solutions that must be explored. It may be possible to use graph-theoretic methods to find optimal rectangularizations (see e.g., Eppstein, 2010), but currently known methods are designed to minimize the number of rectangles rather than some cost function on rectangles.

## SIMPLICITY AND INFORMATIVENESS IN SEMANTIC CATEGORY SYSTEMS 9

Language	Illustrative binary description	Optimal codelength	Complexity
	<p>00000101 00001001</p> <p>01000000 00101000 0010101 00110000</p> <p>0001111 00100001 00101001 00110001</p> <p>00000001</p>	<p>min = 15.81 bits</p> <p>min = 29.29 bits</p> <p>min = 29.29 bits</p> <p>min = 7.91 bits</p>	<p><math>\Sigma = 82.3</math> bits</p>
	<p>00100000 0111100</p> <p>00011001 0110101 00101001</p> <p>100101 00110000</p>	<p>min = 13.81 bits</p> <p>min = 21.14 bits</p> <p>min = 13.64 bits</p>	<p><math>\Sigma = 48.59</math> bits</p>
	<p>1000000</p> <p>1000001</p> <p>100001</p> <p>100010</p>	<p>min = 6.32 bits</p> <p>min = 6.32 bits</p> <p>min = 6.32 bits</p> <p>min = 6.32 bits</p>	<p><math>\Sigma = 25.29</math> bits</p>

*Figure 4.* Computing the complexity of three example languages in a  $4 \times 4$  universe. Categories are indicated by the four colors. Each category is described by a set of rectangle symbols that minimize codelength. The top language does not permit a short description and is therefore considered complex. The middle language consists of three contiguous categories which permit a shorter description, so it is therefore considered less complex. The bottom language, which has a very short description, is the simplest four-category language. The binary strings are concatenated from the codewords shown in Fig. 3.

where  $\gamma > 0$  controls how quickly similarity decays with distance and  $d(\cdot, \cdot)$  gives the distance between two meanings in  $(U, d)$ . In all work reported below, we set  $\gamma = 1$  and  $d$  is the Euclidean metric. This models categories as Gaussians in which the most prototypical meaning (the meaning at the geometric center of the category with the greatest similarity to other category members) has the highest probability of being inferred by a hypothetical listener. This is illustrated in Fig. 5, but note that no interaction is actually played out between our agents; rather, the learner has a bias favoring languages that would hypothetically be more informative in expected communicative scenarios.

Communicative cost predicts three key properties that make a semantic category system informative:

1. **Expressivity** The more categories a system has, the more informative that system is (i.e., communicative cost is lower).
2. **Balanced cardinality** The more evenly-balanced the category sizes are, the more informative the system is (i.e., communicative cost is lower).
3. **Compactness** The more compact categories are, the more informative the system is (i.e., communicative cost is lower).

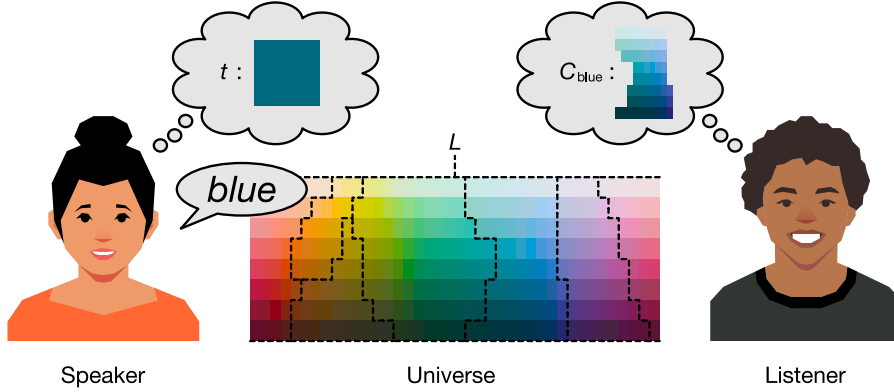


Figure 5. A speaker wishes to communicate a target meaning  $t$  from a universe of meanings. She determines which category the target belongs to according to the shared language  $L$ , which partitions the universe into categories, and utters its associated word. The listener maps this word back to a category of possible meanings and must decide on a specific meaning to infer. He is more likely to infer prototypical (central) members of the category. Informative languages with low communicative cost are structured in such a way as to minimize the potential loss of information that occurs during interaction.

Thus, under the informativeness bias, the learner will prefer systems that exhibit these three properties. Of these, the compactness property is especially important to this paper. By “compactness” we mean the extent to which similar meanings belong to the same semantic category and dissimilar meanings belong to different semantic categories. Sometimes this property is described as “well-formedness” (Regier, Kay, & Khetarpal, 2007) or by the similar notion of “convexity” (Gärdenfors, 2000). Compact categories are informative because they minimize the distance between the speaker’s intended meaning and the listener’s inferred meaning; if a speaker has a particular color in mind and utters the word *blue*, the extent to which the listener will successfully infer the speaker’s intended meaning is a function of how compact the BLUE category is in their shared language. For a more complete introduction to communicative cost, consult the references above or supplementary item S1.

**Posterior.** On observing data  $D$ , a Bayesian agent samples a language  $L$  from the posterior distribution over the space of language hypotheses  $\mathcal{L}$ . The posterior is given by

$$p(L|D; \pi, w, \epsilon) \propto p(D|L; \epsilon) \pi(L)^w, \quad (8)$$

where  $w$  is a free parameter determining the strength of the prior. In all work that follows, we set  $N_{\max} = 4$  (an agent is limited to inferring at most four categories) and we assume an  $8 \times 8$  universe, so the number of language hypotheses is  $|\mathcal{L}| = 4^{64}$ . Since we cannot sample directly from a hypothesis space of this size, we use the Metropolis–Hastings algorithm, which is initialized with a random language  $L_0$ . To select the language at the next step,  $L_{i+1}$ , we propose a candidate language  $L'$  and then calculate the acceptance ratio  $\alpha$ , given by

## SIMPLICITY AND INFORMATIVENESS IN SEMANTIC CATEGORY SYSTEMS 11

$$\alpha = \frac{p(L'|D; \pi, w, \epsilon)}{p(L_i|D; \pi, w, \epsilon)} \cdot \frac{p(L_i|L')}{p(L'|L_i)}. \quad (9)$$

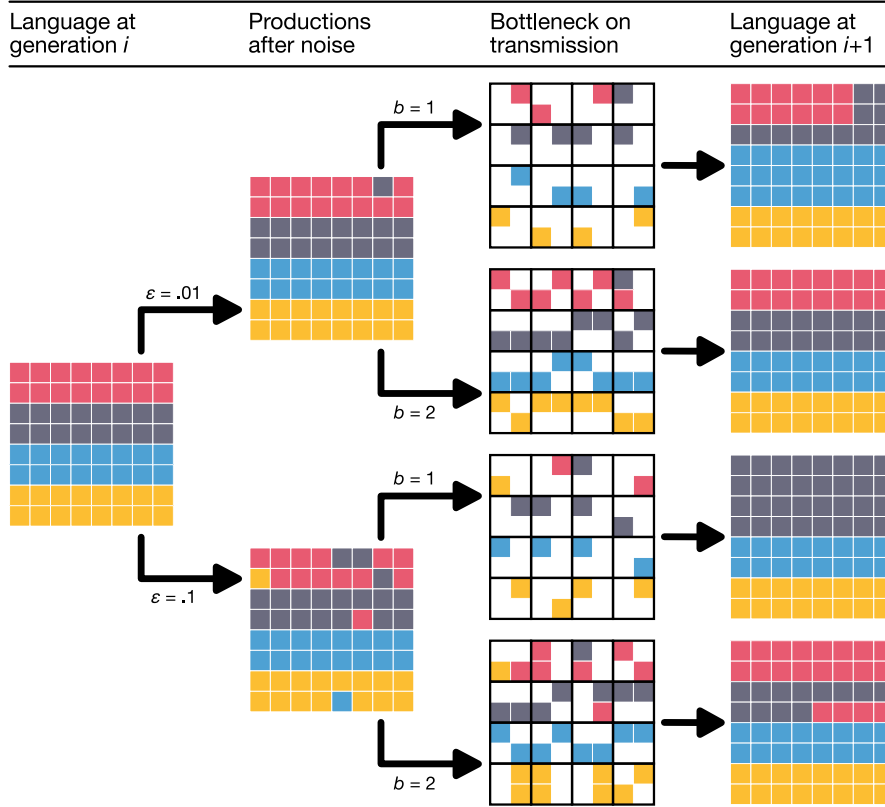
To propose a candidate language, a rectangular region in  $U_{8 \times 8}$  is chosen at random such that all meanings in that region belong to a single category according to  $L_i$ ; these meanings are then transferred to one of the other three possible categories at random, forming the new candidate  $L'$ .<sup>4</sup> This proposal function is asymmetric— $p(L'|L_i) \neq p(L_i|L')$ —which is accounted for by the proposal ratio in Equation 9. Finally, the candidate language is accepted ( $L_{i+1} = L'$ ) if  $\alpha \geq 1$  or with probability  $\alpha$  if  $\alpha < 1$ ; otherwise the candidate is rejected and the previous state is retained ( $L_{i+1} = L_i$ ). This process is repeated 5000 times, and the final state is taken to be a fair sample from the posterior.

**Iterated learning.** Agents are organized into chains such that the production output of one agent becomes the training input to the following agent in the chain, subject to noise on production and the bottleneck on transmission, a limit on how much information is transmitted from one generation to the next. An agent produces signals for each of the 64 meanings in  $U_{8 \times 8}$  according to Equation 1, such that any given signal may be a production error with probability  $\epsilon$ . The meaning–signal pairs that pass through the bottleneck are selected pseudorandomly to ensure that the following agent in the chain sees a uniform spread over the whole space (this becomes more relevant to the experiments reported later). Specifically, the  $8 \times 8$  space is broken into  $16 \times 2 \times 2$  segments and a fixed number of meanings  $b \in \{1, 2, 3, 4\}$  are randomly selected (without replacement) from each segment,  $b$  being the bottleneck parameter (see Fig. 6 for examples). Finally, we also consider the exposure level  $\xi$  which controls how many exposures an agent gets to the dataset (i.e., each meaning–signal pair that passes through the bottleneck is observed  $\xi$  times).

**Summary.** The model provides five free parameters (summarized below), which we manipulate and discuss in the following section.

1. **Prior ( $\pi$ )** We consider two prior functions as models of two extreme positions. First, a prior for simplicity,  $\pi_{\text{sim}}$ , motivated by the principle of Occam’s razor in inductive reasoning. Second, a prior for informativeness,  $\pi_{\text{inf}}$ , motivated by the theory that learners are biased toward making languages more informative.
2. **Weight ( $w$ )** The weight parameter affects the strength of the prior bias. When  $w = 1$  the prior is left unchanged; when  $w > 1$ , the prior is strengthened; when  $0 < w < 1$  the prior is weakened; and when  $w = 0$  the prior is flattened to a uniform distribution.
3. **Bottleneck ( $b$ )** The size of the bottleneck determines how much data passes from one generation to the next (specifically,  $b$  is the number of meanings selected from each  $2 \times 2$  segment).
4. **Exposures ( $\xi$ )** The number of exposures determines how many times an agent is exposed to the dataset  $D$ , the meaning–signal pairs that passed through the bottleneck.

<sup>4</sup>A simpler symmetric proposal function in which single meanings are moved between categories at each step is prone to getting stuck in local maxima, which limits the ability of the algorithm to freely explore the hypothesis space under either prior function. Note that this method is not biased toward introducing new rectangles because it only modifies the category membership of rectangular areas that already exist.



*Figure 6.* Illustration of the transmission procedure over a single generation. The agent at generation  $i$  has a language (column 1), which it uses to produce signals with some probability of noise  $\epsilon$  (column 2). These productions are passed through the bottleneck on transmission by dividing the universe into 16  $2 \times 2$  segments (black grids) and selecting  $b$  meanings from each segment (column 3). The agent at generation  $i + 1$  only sees the signals associated with these meanings and, aided by the prior, must generalize to unseen meanings (white cells) forming a new language (column 4). The language is transmitted most faithfully when there is low noise (e.g.,  $\epsilon = .01$ ) and a wide bottleneck (e.g.,  $b = 2$ ).

## SIMPLICITY AND INFORMATIVENESS IN SEMANTIC CATEGORY SYSTEMS 13

5. **Noise ( $\epsilon$ )** The noise parameter affects how many production errors an agent makes when producing signals for the following generation to observe. It also appears in the likelihood function, encoding an expectation of how much noise there is in the data.

### Results

Results are shown in Fig. 7 under the parameter settings  $b = 2$ ,  $\xi = 2$ , and  $\epsilon = .01$  (for the full set of model results under 48 parameter combinations, see supplementary item S2 or <https://joncarr.net/p/shepard/>). We contrast three prior biases: the simplicity prior with  $w = 1$ , the informativeness prior with  $w = 1$ , and a strong form of the informativeness prior with  $w = 500$ . This strong form of the informativeness prior emphasizes the compactness property discussed earlier; agents with this prior bias have a strong proclivity toward compact categories that minimize the potential for communicative error. We consider four quantities of interest: expressivity (the number of categories inferred), complexity (Equation 4), communicative cost (Equation 6), and transmission error, which is measured as the variation of information (VI; see Meilă, 2007) between the language in a particular generation ( $L$ ) and the language in the previous generation ( $L'$ ):

$$\text{VI}(L, L') = - \sum_{C \in L, C' \in L'} \frac{|C \cap C'|}{|U|} (\log \frac{|C \cap C'|/|U|}{|C|/|U|} + \log \frac{|C \cap C'|/|U|}{|C'|/|U|}). \quad (10)$$

Under this measure, an agent who fully reproduces the partition used by the previous agent in the chain will have VI of 0 bits; maximum VI is  $\log |U| = 6$  bits.

The results under the simplicity prior are shown by the blue lines in Fig. 7 and a typical chain is depicted in Fig. 8A. Over 50 generations, the languages become less complex, which is achieved in two ways: First, the categories take on simple, contiguous structures that may be described by a shorter description in the rectangle code; and second, categories are gradually lost over time, further simplifying the languages. This has an interesting effect on communicative cost, which initially drops—implying more informative languages—but then begins to rise again. This is because the contiguous categories that initially emerge are generally quite compact, and since communicative cost is sensitive to compactness, it initially decreases. But this effect is then gradually eroded by the loss of expressivity. Furthermore, the category structures that arise under the simplicity prior tend to mark distinctions on just one of the two dimensions; in the example in Fig. 8A, the language ends up marking a three-way distinction on the  $x$ -axis. This overall process of simplification results in more learnable languages, as indicated by decreasing transmission error over time. Within around 10 generations, the languages have simplified into configurations that are reliably transmitted from one generation to the next, despite the fact that agents only receive input data for half ( $b = 2$ ) of the meanings.

The results under the basic informativeness prior ( $w = 1$ ) are shown by the solid red lines in Fig. 7 (see Fig. 8B for a typical example). The bias for informativeness causes the agents to maintain all four categories in well-balanced proportions, but there is no effect on transmission error, complexity, or communicative cost. This is because the prior is very flat with respect to compactness and mostly encodes a preference for greater expressivity, which cannot be obtained because of the  $N_{\max} = 4$  limit that we have imposed. If we

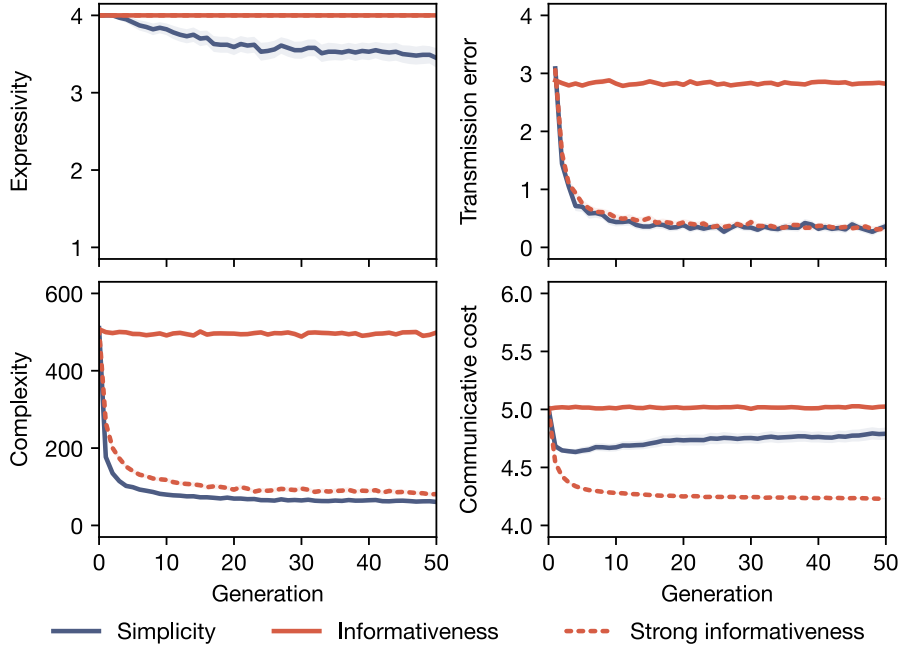


Figure 7. Results for expressivity, transmission error, complexity, and communicative cost under a simplicity prior ( $w = 1$ , solid blue), informativeness prior ( $w = 1$ , solid red), and a strong form of the informativeness prior ( $w = 500$ , dashed red). Plots show the mean of 100 chains over 50 generations, and the shaded areas show the 95% confidence intervals on the mean. Agents see half of the meanings ( $b = 2$ ) in two exposures ( $\xi = 2$ ) with a 1% chance of noise on production ( $\epsilon = .01$ ).

were to remove this limit, expressivity would rise to 64 categories (every meaning forms its own category), communicative cost would decrease to its minimum value of 0 bits, and complexity would increase to its maximum value of around 715 bits.

The dashed red lines in Fig. 7 show results under the strong informativeness prior ( $w = 500$ ; see Fig. 8C for an example). When the informativeness prior is strengthened in this way, all four categories continue to be maintained in well-balanced proportions, but communicative cost also experiences a sustained decrease because the stronger prior favors categories that are maximally compact. This pressure for compactness drives chains toward one special partition of the universe: the partition into quadrants, as seen in the final generation in Fig. 8C. This quadrant partition is the optimal packing of the space into four equally-sized categories. Within a category, the Euclidean distance is minimized between any two category members; such a system is informative because it minimizes the potential degree of communicative error.

The extent to which chains converge to the prior bias—be it for simplicity or informativeness—is essentially controlled by intergenerational information loss, which Spike,

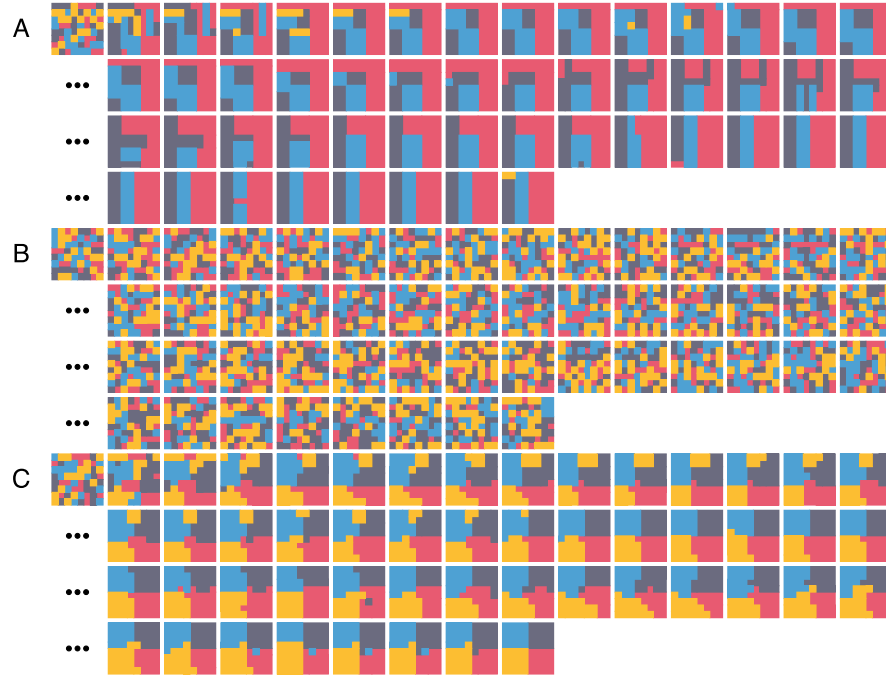


Figure 8. An example chain for each of the three prior biases: (A) the simplicity prior, (B) the informativeness prior, and (C) the strong informativeness prior. Each chain consists of the randomly generated language that initialized the chain (generation 0) followed by 50 generations of iterated learning. Each chain is broken across four rows.

Stadler, Kirby, and Smith (2017) argue is an essential requirement in the emergence of structured languages. The greater the information loss, the faster the convergence to the prior. In other words, the less data an agent has to rely on, or the more unreliable that data is, the more the agent must lean on its prior bias to reconstruct the language. This is illustrated in Fig. 9, which shows the distributions of complexity scores in the final (50th) generation under the simplicity bias. There is greater convergence to the prior bias for simplicity when the bottleneck is tighter, there are fewer exposures, or the level of noise is greater.

### Summary

We have put forward a model of a Bayesian language learner and have considered two prior functions: one for simplicity and one for informativeness. These two priors represent two extreme positions that one may take in regard to the learning of semantic categories. In addition, we consider what happens when the informativeness prior is strengthened such that the compactness component of informativeness is magnified. The results show that, when the number of categories is limited to four (i.e.,  $N_{\max} = 4$ ), as is the case in Carstensen



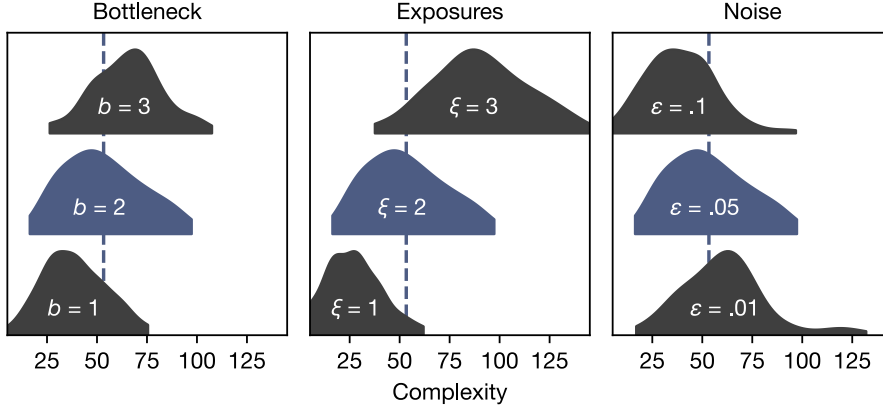


Figure 9. Each density plot shows how complexity is distributed in the final-generation languages across 100 chains. The distributions in blue are identical and show results under the simplicity bias with parameter settings  $b = 2$ ,  $\xi = 2$ , and  $\epsilon = .05$ ; dashed lines show mean complexity under these parameters. The distributions in black show what happens when one parameter is manipulated, while holding the other two constant. Chains become simpler faster when there is a tighter bottleneck, fewer exposures, or noisier productions. Distributions are scaled to equal height to highlight their shape.

et al. (2015, Study 2), communicative cost only decreases in two situations. Either agents must have a simplicity bias, in which case the decrease in communicative cost is predicted not to be sustained in the long run, or agents must have a strong form of the informativeness bias (i.e., a form of the informativeness bias that strongly favors compact categories because they minimize communicative error). If the simplicity prior offers a better model, we would expect to find that iterated learning results in a loss of expressivity and leads to contiguous category structures that make distinctions on principally one dimension (because these types of system are simpler). If the informativeness prior offers a better model, we would expect to find that iterated learning maintains high expressivity and—if the compactness component of the bias is especially strong—it should lead to a sustained decrease in communicative cost and maximally compact category structures, the optimal configuration being a partition of the universe into quadrants. We test these predictions in two experiments.

### Experiment 1

In Experiment 1, we test for a difference in learnability between two basic types of category structure—stripes and quadrants—which are illustrated in Fig. 10 (specific details about the stimuli will be explained shortly). The model presented above yields predictions about how easy these systems should be to learn. If learners have an inductive bias for simplicity, then simple category structures (i.e., stripes) should be learned more easily than more complex structures (i.e., quadrants). If learners have an inductive bias for informativeness then the model makes two predictions: Under the basic informativeness

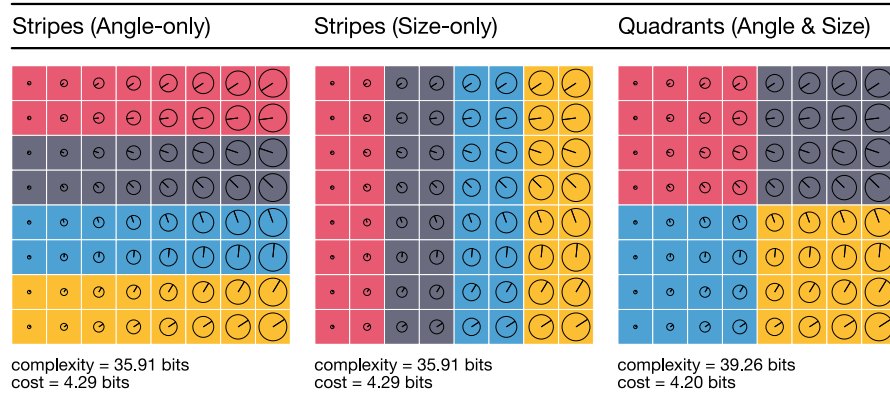


Figure 10. The three category systems in Experiment 1. The stimulus space is discretized into four equally-sized categories in three ways. The Angle-only and Size-only partitions are simple. The Angle & Size partition is more complex but offers a potential benefit in terms of informativeness because category members are more tightly packed.

Table 2

Model predictions tested by Experiment 1

Inductive bias	Prediction made by the model
Simplicity	Stripes easier to learn than quadrants (because stripes are simpler)
Informativeness	Stripes and quadrants are equally easy to learn (because both are approximately equally informative)
Strong informativeness	Quadrants easier to learn than stripes (because quadrants are more informative/compact)

bias ( $w = 1$ ), stripes and quadrants should be similarly learnable because they are (approximately) equally informative (four categories each with 16 members); under stronger forms of the informativeness bias ( $w > 1$ ), quadrants should be easier to learn than stripes because they constitute a more compact packing of the space, minimizing the potential for communicative error. These predictions are summarized in Table 2 and illustrated more formally in Fig. 11.

## Method

The experiment was a simple category learning task. Participants were first trained on one of three category systems (approximately 15 min) and were then tested to see how well they learned the system (approximately 5 min). We test participants through either a production test (given a stimulus, supply a label) or a comprehension test (given a label, supply a stimulus). The production version of the experiment is matched to the model; the comprehension version was included in case the effect of an informativeness bias in learning only manifests itself when participants recognize that they need to be able to comprehend

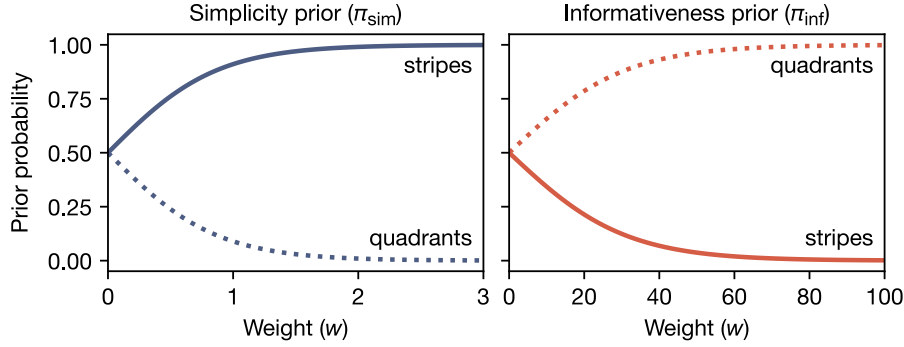


Figure 11. Assume that an agent must choose between two hypotheses, a striped partition or a quadrant partition, given data that is equally likely under either. The simplicity and informativeness priors make opposite predictions about which hypothesis will be inferred, which is exaggerated as the strength of the bias is increased. If the prior is uniform ( $w = 0$ ), the agent has no prior preference between the two languages.

the language. The experiment therefore has a  $3 \times 2$  design: Participants were assigned randomly to one of three category systems (as detailed below) and to one of two test types (Production or Comprehension).

**Participants.** 240 participants were recruited through the Figure Eight platform (<https://www.figure-eight.com>), 40 in each of the six conditions.<sup>5</sup> Participants were paid \$3.00 for participation, plus up to \$1.92 in bonuses based on the accuracy of their learning (as detailed below). Ethical approval was granted according to the procedures of the School of Philosophy, Psychology and Language Sciences at the University of Edinburgh. All participants provided informed consent.

**Stimuli.** We adopted the so-called “Shepard circles” (Shepard, 1964) as our stimulus space (see Fig. 10), which vary continuously in angle and size. The space was quantized onto an  $8 \times 8$  grid, yielding 64 discrete stimuli. The radii increase linearly from 25 pixels to 200 pixels and the angles increase linearly over  $180^\circ$  from  $147^\circ$  to  $327^\circ$ . The stimuli closely replicate the Shepard circles used by Canini et al. (2014, Fig. 1, p. 787), who showed in a multidimensional scaling analysis that participants’ dissimilarity perceptions of these stimuli are closely correlated with the Euclidean distance in the  $8 \times 8$  grid. This makes the stimuli well justified analogs of the abstract meanings used in the model and allows us to assume that the Euclidean distance in the  $8 \times 8$  grid is an acceptable approximation of perceived dissimilarity.

The three category systems differ in how the stimulus space is partitioned into four categories, as shown in Fig. 10. In the Angle-only system, the categories mark a four-way distinction in angle; in the Size-only system, the categories mark a four-way distinction in

<sup>5</sup>A total of 309 participants began the experiment, but three were excluded because they repeatedly clicked the same response button, and a further 66 terminated the experiment prior to completing it, so their data were erased because they were deemed to have withdrawn consent. See supplementary item S3 for additional details.

## SIMPLICITY AND INFORMATIVENESS IN SEMANTIC CATEGORY SYSTEMS 19

Table 3  
*Category labels*

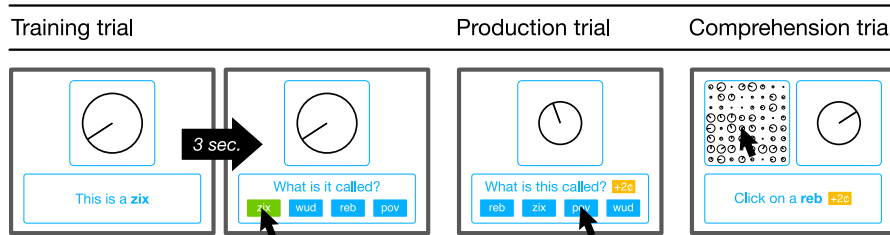
Set	Labels			
1	<i>pov</i>	<i>reb</i>	<i>wud</i>	<i>zix</i>
2	<i>gex</i>	<i>juf</i>	<i>vib</i>	<i>wop</i>
3	<i>buw</i>	<i>jef</i>	<i>pid</i>	<i>zor</i>
4	<i>fod</i>	<i>jes</i>	<i>wix</i>	<i>zuv</i>

size; and in the Angle & Size system, the categories mark a four-way distinction that relies on both dimensions. In all three systems, the number of categories (four) and the cardinality of categories (16) is equal, but they still differ in terms of simplicity and informativeness. The striped category systems (Angle-only and Size-only) are simpler (complexity = 35.91 bits), since only one dimension needs to be considered; in contrast, the quadrant system (Angle & Size) is more complex (complexity = 39.26 bits), since both dimensions need to be considered. However, the more compact packing of the Angle & Size system makes it marginally more informative (cost = 4.20 bits) than the other two category structures (cost = 4.29 bits), a difference that is magnified if we assume a stronger form of the informativeness prior.

The category labels used to label the stimuli were three-letter nonwords (see Table 3). To create the labels, we generated all CVC strings, such that the first consonant letter was not the same as the final one, and then removed valid English words (e.g., *pin*). For each of the remaining candidate labels, we attempted to translate the word into English from each of the 63 languages that use the Latin script in Google Translate. If Google was unable to offer a translation, we assumed that the label was not a real word in that language. We then selected 16 unique labels that were meaningless in as many languages as possible, such that they could be arranged into four sets of four labels in which any given set used no letter more than once. Each participant was assigned one set of labels selected at random, and the mapping between labels and categories was randomized for every participant. This procedure was designed to mitigate possible interference from native language and to reduce the possibility of iconic meaning-signal correspondences occurring, potentially making some mappings easier to learn than others (see e.g., Nielsen & Rendall, 2012; Nygaard, Cook, & Namy, 2009).

**Training procedure.** In the training phase, participants were trained on half of the 64 stimuli. These 32 stimuli were selected pseudorandomly through the same bottlenecking procedure used in the model (see Fig. 6), which ensures that an equal number of training items (eight) are selected from each category. Training on the 32 items was repeated four times (i.e., in four blocks), since initial piloting indicated that participants would need at least four exposures to perform well above chance.

In each training block, the participant was exposed to each of the 32 training items in random order. In a single exposure (see Fig. 12), the stimulus was presented first, and after a one-second delay, the sentence “This is a **zix**” appeared containing the relevant category label; this sentence remained on screen for 3 s, at which point it was replaced by the question “What is it called?” along with four buttons showing the four possible labels (the order of the label buttons was randomized on every trial). If the participant clicked the correct button, the button turned green; if incorrect, the button turned red and the button



*Figure 12.* Illustration of training and test trials (not to scale). In training trials, the participant is shown a stimulus and its category label; after 3 s they must click the correct label. Participants cannot simply rely on short-term memory because training is also interspersed with mini-tests in which participants are tested on a previously-seen item. Test trials (and mini-test trials) differ according to condition: The participant is either given a stimulus and must supply a label (Production) or given a label and must supply a stimulus (Comprehension). A 2c bonus is awarded for every correct test or mini-test response.

for the correct label turned green. After every fourth training trial (i.e., eight times per block, 32 times in total), a “mini-test” was inserted. Mini-test trials took the same form as the test trials according to condition as described below. The participant was awarded 2c for every correct mini-test response; a running total at the top of the screen recorded the total bonuses earned during training. The purpose of the mini-tests and bonusing scheme was to keep participants interested and highly incentivized to learn the language as well as possible.

**Production procedure.** The test phase differed according to condition. Participants assigned to the Production test-type were asked to label all 64 stimuli. On each test trial, a stimulus was presented alongside the question “What is this called?” (see Fig. 12). After a 1 s delay, the set of four labels appeared below in random order. It was made clear to the participant that they would be awarded 2c for every correct response; however, no feedback was given during the test, including no running total of the bonuses earned. The lack of feedback during the test was designed to elicit responses based on the hypothesis fixed during training.

**Comprehension procedure.** For participants assigned to the Comprehension test-type, mini-test and test trials took a different form (see Fig. 12). In these trials, the participant was presented with a sentence like “Click on a **reb**” and, after a 1 s delay, was provided with an  $8 \times 8$  grid of thumbnail images showing all 64 stimuli in random order. This was accompanied by a stimulus viewer which showed the stimuli at full size as the participant hovered their cursor over the thumbnails. Like Production, there were 64 trials, 16 per label, with trials in random order. The participant was awarded a 2c bonus for every correct answer; any of the 16 items from the relevant category was considered correct, so the payoff structure is identical under either test type—in both cases there is a  $1/4$  probability of a correct response by chance.

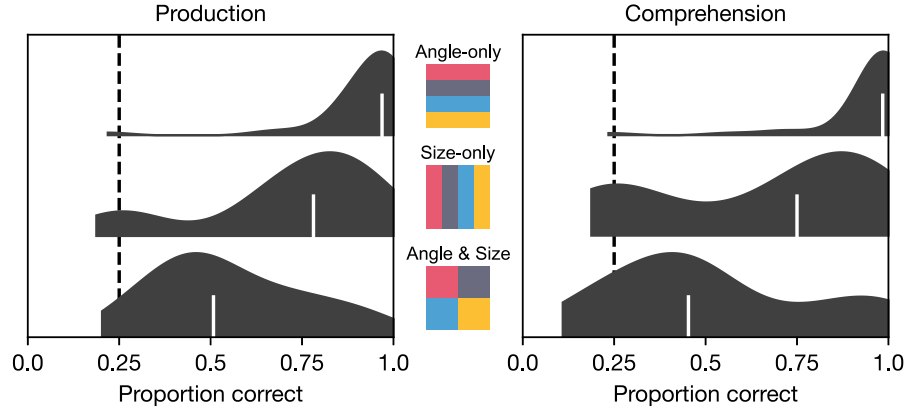


Figure 13. Density plots showing the distribution of participants' test accuracies (proportion of test responses that were correct) by condition. White lines show the medians; the dashed lines show chance level. Participants learning the Angle-only system had the highest accuracy and participants learning the Angle & Size system had the lowest accuracy, regardless of whether participants were tested through production or comprehension. Distributions are scaled to equal height to highlight their shape.

## Results

Fig. 13 shows how accurate participants were by condition (the proportion of test responses that were correct). A linear mixed-effects regression analysis was used to test for an effect of condition on a participant's chance of giving a correct response on a given test trial with participant as a random effect.<sup>6</sup> Helmert contrast coding was used to test the category systems in a step-wise fashion (i.e., first whether the Angle-only and Size-only systems differ, and then whether those systems combined differ from the Angle & Size system). There was no significant difference between Production and Comprehension ( $\beta = 0.20 \pm 0.23$ ,  $p = .371$ ). The Size-only system was significantly harder to learn than the Angle-only system ( $\beta = -1.10 \pm 0.20$ ,  $p < .001$ ). Crucially, and as hypothesized, participants found the Angle & Size system significantly more difficult to learn than the striped systems ( $\beta = -0.66 \pm 0.11$ ,  $p < .001$ ). For illustrations of individual participant results, see supplementary item S4.

## Summary

Experiment 1 directly addresses a predicted difference between learning biases for simplicity and informativeness. If learners have an informativeness bias that strongly prefers compact categories, we would expect to find that the quadrant configuration is readily inferred, but this was not the case. Participants found the quadrant configuration difficult

<sup>6</sup>All statistical analyses reported in this paper were conducted using version 1.1.13 of the R package lme4 (Bates, Mächler, Bolker, & Walker, 2015).

## SIMPLICITY AND INFORMATIVENESS IN SEMANTIC CATEGORY SYSTEMS 22

Table 4

*Model predictions tested by Experiment 2*

Inductive bias	Predictions made by the model
Simplicity	Expressivity decreases over generations
	Transmission error decreases over generations
	Complexity decreases over generations
	Cost decreases and then increases over generations
Informativeness	Expressivity is maintained over generations
	Transmission error is maintained over generations
	Complexity is maintained over generations
	Cost is maintained over generations
Strong informativeness	Expressivity is maintained over generations
	Transmission error decreases over generations
	Complexity decreases over generations
	Cost decreases over generations

to learn, while the striped partitions were significantly easier, as predicted under a simplicity bias.

## Experiment 2

In Experiment 2, we conducted an iterated learning experiment closely matched to the model described earlier in this paper. The aim of this experiment was to test whether informative systems could emerge through iterated learning, as found by Carstensen et al. (2015). Perhaps, for example, a bias for informativeness only manifests itself when amplified by iterated learning. Table 2 summarizes the predictions made by the model in terms of what should happen under different inductive biases in iterated learning (see also the model results in Fig. 7). A simplicity bias predicts a generational decrease in expressivity, transmission error, and complexity, and an initial decrease in communicative cost that would not be sustained over subsequent generations. An informativeness bias predicts that expressivity will be maintained over generations and, if the bias strongly prefers compactness, a sustained decrease in communicative cost, which also leads to a decrease in complexity and transmission error.

## Method

From the point of view of the participant, Experiment 2 was identical to the Production version of Experiment 1. Participants learn and reproduce labels for stimuli, but—unknown to them—their production output is passed on to a new participant, whose production output is in turn passed on to another new participant, following a standard iterated learning design.

## SIMPLICITY AND INFORMATIVENESS IN SEMANTIC CATEGORY SYSTEMS 23

**Participants.** 224 participants were recruited through the Figure Eight platform.<sup>7</sup> Participants who had already taken part in Experiment 1 were not able to take part in Experiment 2. Payment was identical to Experiment 1. Ethical approval was granted according to the procedures of the School of Philosophy, Psychology and Language Sciences at the University of Edinburgh. All participants provided informed consent.

**Transmission procedure.** As in the model, the initial participant in a chain was given a randomly generated language to learn with an equal number of meanings in each category. The language they produced during the test phase was then transmitted to a new participant, subject to the same bottlenecking procedure as the model ( $b = 2$ ). Participants were assigned to one of 12 chains at random, and chains were run for a minimum of 10 generations. After the 10th generation, we allowed the chains to continue running until they eventually converged on a particular categorization system. Chains were deemed to have converged when two consecutive participants infer exactly the same language, suggesting that that language is especially easy to learn—an attractor in the space of possible languages.

## Results

Fig. 14 depicts the evolution of all 12 chains (labeled A–L) through to convergence, and the converged-on languages are summarized in Fig. 15. In two cases, the languages collapsed to a single category; in one case, a two-category system emerged; and in eight cases, a three-category system emerged. Only in one case did the language retain all four categories. Fig. 15 also highlights the category structures that emerged. In eight cases (dashed black box), a system of contiguous categories emerged marking distinctions on the angle dimension, while in two cases (dashed gray box), a system of contiguous categories emerged marking distinctions on the size dimension. This is consistent with Experiment 1 where we found that size-based systems were harder to learn than angle-based systems; the languages are adapting to the learning biases of their learners. These findings speak directly to the predictions made by the simplicity prior in the model: We observe a loss of expressivity and the emergence of simple, contiguous category structures that make distinctions on principally one dimension.

Fig. 16 shows the results under the same four quantitative measures computed for the model. Results are shown only for the first 10 generations, for which we have data from all 12 chains. A linear mixed-effects regression analysis was used to test for an effect of generation on each of the four measures with chain as a random effect and by-chain random slopes for the effect of generation (following the procedure recommended by Winter & Wieling, 2016).  $P$ -values were obtained by likelihood ratio tests of the full model against a null model without the effect in question. To be conservative, generation 0 is not included because it is not derived from participant data. As predicted by the simplicity prior in the model, expressivity ( $\beta = -0.06 \pm 0.03$ ,  $\chi^2 = 5.19$ ,  $p = .023$ ), transmission error ( $\beta = -0.22 \pm 0.05$ ,  $\chi^2 = 13.22$ ,  $p < .001$ ), and complexity ( $\beta = -22.59 \pm 6.12$ ,  $\chi^2 = 9.67$ ,  $p = .002$ ) all decreased with generation. Over time, the languages become simpler and less expressive,

<sup>7</sup>A total of 273 participants began the experiment, but one was excluded because they repeatedly clicked the same response button, and a further 48 terminated the experiment prior to completing it, so their data were erased because they were deemed to have withdrawn consent. See supplementary item S3 for additional details.



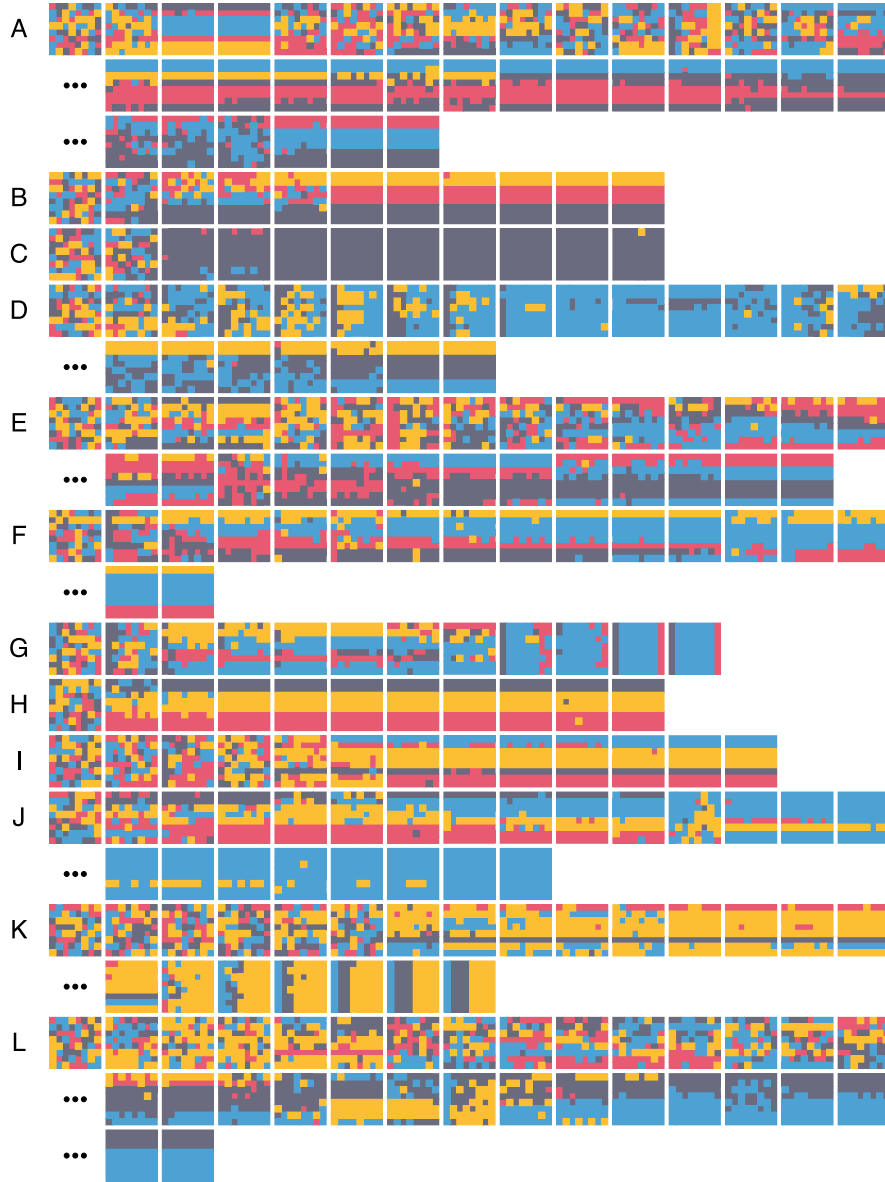


Figure 14. Evolution of all 12 chains (labeled A–L) in Experiment 2. Each chain consists of the randomly generated language that initialized the chain (generation 0) followed by the language produced by each subsequent participant in the chain. Some chains are broken across multiple rows. Chains were run for a minimum of 10 generations, after which they continued to run until two consecutive participants inferred exactly the same language.

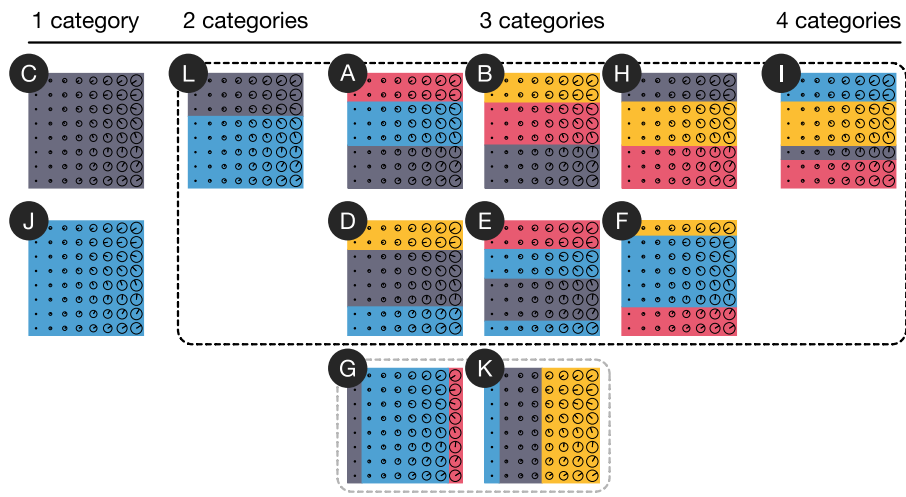


Figure 15. The languages that were converged on across all 12 chains (labeled A–L) grouped by number of categories. Eight of the emergent languages mark distinctions in angle (dashed black box) and two of the languages mark distinctions in size (dashed gray box).

and as a result more faithfully transmitted. Contrary to Carstensen et al. (2015), we did not find a decrease in communicative cost over generations ( $\beta = 0.009 \pm 0.01$ ,  $\chi^2 = 0.46$ ,  $p = .499$ ), implying that the languages are *not* becoming more informative.

### Model fit

Although the experimental results reported above are strongly suggestive that the simplicity prior offers a better model of semantic category learning, we can also estimate the parameters of our model from the experimental data to objectively determine which prior function offers the better fit. The experiment fixes two parameters, the bottleneck  $b = 2$  and the number of exposures  $\xi = 4$  (i.e., participants get four exposures in four training blocks), while the prior function  $\pi$ , its weight  $w$ , and the noise level  $\epsilon$  are unknown. To estimate these parameters from the experimental data, we start by defining the maximum a posteriori (MAP) probability that a participant would produce certain output data ( $D_{\text{out}}$ ) given their input data ( $D_{\text{in}}$ ) under certain parameter assumptions:

$$p_{\text{MAP}}(D_{\text{out}}|D_{\text{in}}; \pi, w, \epsilon) = \max_{L \in \mathcal{L}} \frac{p(D_{\text{out}}|L; \epsilon)p(L|D_{\text{in}}; \pi, w, \epsilon)}{\sum_{L' \in \mathcal{L}} p(D_{\text{out}}|L'; \epsilon)p(L'|D_{\text{in}}; \pi, w, \epsilon)}, \quad (11)$$

where  $p(D_{\text{out}}|L; \epsilon)$  is the likelihood from Equation 2, and  $p(L|D_{\text{in}}; \pi, w, \epsilon)$  is the posterior from Equation 8, which now becomes a prior. By maximizing over the space of possible languages, this probability is defined in terms of the language that the participant is most likely to have had in mind given the data they received and the data they produced (under certain parameter assumptions).

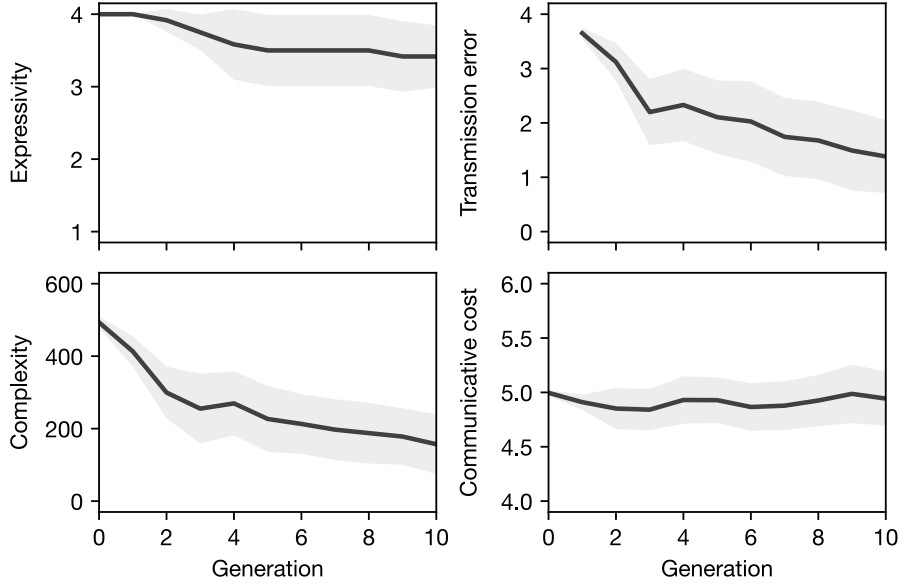


Figure 16. Experimental results for expressivity, transmission error, complexity, and communicative cost. Over 10 generations, the languages become simpler, easier to learn, and less expressive. Communicative cost remains static implying that the languages are not becoming more informative.

Guided by Equation 11, we estimate the likelihood of our experimental dataset  $\mathcal{D}$  (i.e., participants'  $\langle D_{\text{in}}, D_{\text{out}} \rangle$  pairs<sup>8</sup>) under settings of  $\pi$ ,  $w$ , and  $\epsilon$  as

$$p(\mathcal{D}; \pi, w, \epsilon) = \prod_{\langle D_{\text{in}}, D_{\text{out}} \rangle \in \mathcal{D}} p(D_{\text{out}} | L^*; \epsilon), \quad (12)$$

where  $L^*$  is a language sampled from the posterior  $p(L | D_{\text{in}}; \pi, w, \epsilon)$  using the same techniques applied to the model. In other words, for each participant, we simulate what happens when an agent learns from the participant's  $D_{\text{in}}$ ; the agent infers a language  $L^*$  and we calculate the probability of the agent producing the participant's  $D_{\text{out}}$  given  $L^*$ . We may then seek parameter values that maximize this probability across the experimental dataset as a whole (i.e., parameter values that maximize the probability of an agent producing the same output as a participant when given the same input as that participant). For each of the prior functions,  $\pi_{\text{sim}}$  and  $\pi_{\text{inf}}$ , we estimate  $w^*$  and  $\epsilon^*$ , the parameter values that maximize Equation 12:

<sup>8</sup>The model fit was performed on data from 168 of the 224 participants (75%). We excluded participants whose transmission error was greater than 3 bits, since including them led to very high estimates of noise and a poor fit under either of the priors. In other words, for the purpose of estimating the best parameter values, we retain participants who infer a hypothesis based closely on their input data.

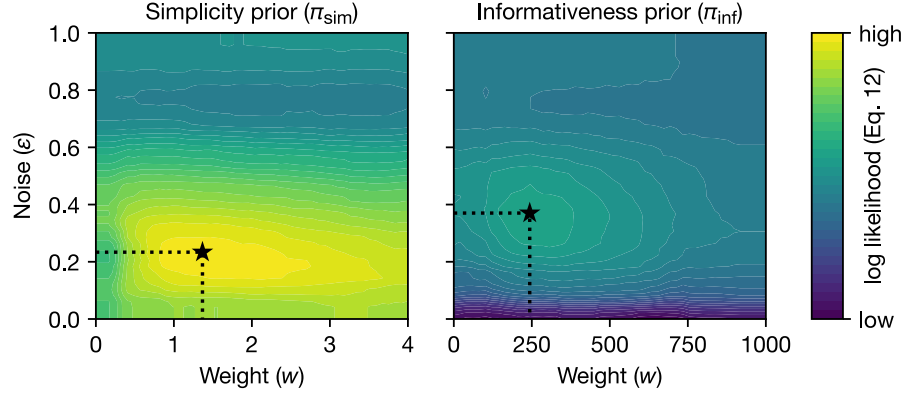


Figure 17. Model fit results for the simplicity prior (left) and informativeness prior (right). Each plot shows how the weight and noise parameters affect the likelihood of observing the experimental dataset. Yellow areas indicate settings of  $w$  and  $\epsilon$  that offer a good fit to the experimental data. The black stars show the maximum likelihood estimates, the parameter values that maximize Equation 12.

$$w^*, \epsilon^* = \arg \max_{w, \epsilon} p(\mathcal{D}; \pi, w, \epsilon). \quad (13)$$

These values were obtained by maximum likelihood estimation using a Bayesian optimizer (Head et al., 2018).  $\epsilon$  was bounded in  $(0, 1)$ .  $w$  was bounded in  $[0, 4]$  for the simplicity prior and  $[0, 1000]$  for the informativeness prior. The upper bounds were selected based on initial experimentation, which indicated that the likelihood would drop off beyond these values.

The results of the model fit are shown in Fig. 17. For the simplicity prior, the maximum likelihood estimates are  $w_{\text{sim}}^* \approx 1.37$  and  $\epsilon_{\text{sim}}^* \approx .23$ , yielding a log likelihood of  $-11323.09$ . For the informativeness prior, the maximum likelihood estimates are  $w_{\text{inf}}^* \approx 243.3$  and  $\epsilon_{\text{inf}}^* \approx .37$ , yielding a log likelihood of  $-17283.35$ . These results tell us that, overall, the best fit to the experimental data is given by a slightly strengthened simplicity prior with a noise level of around 23%. For the informativeness prior, the best fit is obtained by strengthening it and assuming a noise level of around 37%. The likelihood ratio is  $2^{5960}$ , offering overwhelming evidence that the simplicity prior gives a better fit to the experimental data than the informativeness prior.

Rerunning the iterated learning model with the parameter settings  $b = 2$ ,  $\xi = 4$  (matched to the experiment),  $w = w^*$  and  $\epsilon = \epsilon^*$  (estimates from the experimental data), we obtain the results shown in Fig. 18. The experimental results are shown for comparison: Results for the first 10 generations are reproduced from Fig. 16 and results for the subsequent 40 generations (dashed line) are estimated based on the assumption that once a chain fully converges it will experience no further change.<sup>9</sup> These plots confirm that the simplicity prior

<sup>9</sup>In fact, we would expect to see a further reduction in complexity in the experimental results were it financially practical to run all 12 chains for 50 generations. Nevertheless, the assumption that chains are

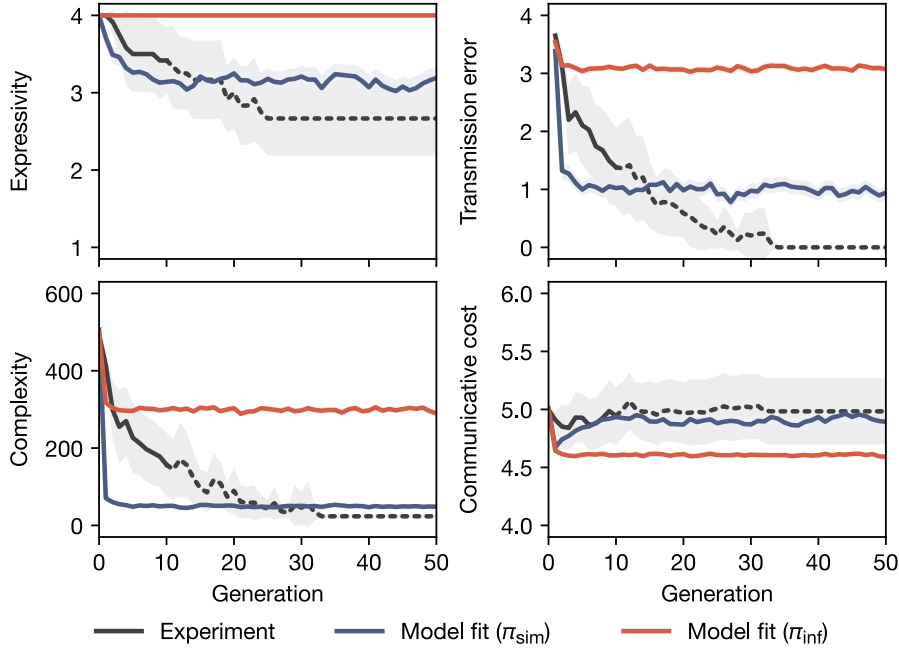


Figure 18. Model results under the simplicity (blue) and informativeness (red) prior using parameters matched to and estimated from the experimental data. For comparison, the experimental results are shown in black. Shaded areas show the 95% confidence intervals on the mean.

yields a general pattern of results that correspond closely to the experimental evidence.

However, Fig. 18 also reveals two discrepancies between the experiment and the model under the simplicity prior: First, the simplicity prior results in a rapid, early decrease in complexity (unlike the experiment), and second, it results in high transmission error later in the chains (unlike the experiment). These discrepancies are a result of our simplistic model of noise on production: The model considers noise to be constant over generations and does not capture the fact that some types of error are more likely than others. In the experiment, however, it appears that noise on production is some function of how complex a language is or how much confidence a participant has in their hypothesis, and participants are also more likely to make certain errors over others (e.g., greater confusion at category boundaries). These aspects of the true dynamics are not captured by our model and could be the subject of future work.

unlikely to change once they converge is not entirely unwarranted: As can be seen in Chains B, C, and H (see Fig. 14), once a chain fully converges, the category structure tends to be reliably conserved over subsequent generations.

### Summary

Experiment 2 shows that, when there is only a pressure from learning, languages evolve to become as simple as possible. Simplicity is achieved by reducing expressivity and moving toward contiguous, one-dimensional categories that have a short description. These results are closely aligned with the simplicity prior in the model; indeed, fitting the model to the experimental data shows that the results are much better predicted by the simplicity prior. This is because, to obtain a good fit under the informativeness prior, it must be strengthened considerably, but when strengthened, the informativeness prior favors a fairly strict four-category quadrant partition rather than fewer striped categories.

### Discussion

Our model and experiments show that the iterated learning of semantic categories—under a theoretically well motivated simplicity prior—leads to simple, relatively uninformative systems. On first glance, this result appears to run contrary to our general experience of the world: Languages are in fact rather informative. We argue that because our model and experiment only include one of the two principal pressures—induction without interaction—the languages are shaped only by a pressure for simplicity. Real languages are informative because there is also some functional reason for them to be so. As such, we expect that by introducing a pressure from interaction (such as through a shared communicative task), the languages will find an optimal tradeoff between simplicity and informativeness. Indeed, this is precisely what is shown in Kirby et al. (2008, 2015) and also in experiments without generational turnover that nevertheless still have pressures from both induction and interaction (Raviv, Meyer, & Lev-Ari, 2018; Winters, Kirby, & Smith, 2018).

Nevertheless, our account still runs contrary to the empirical findings of Carstensen et al. (2015), who found that languages become more informative under iterated learning. Their result appears particularly robust given that it was demonstrated in two studies: Study 1 was a reanalysis of an iterated learning experiment by J. Xu et al. (2013) and Study 2 was a novel iterated learning experiment. What explains the gap between their results and ours? We argue that languages that have the *appearance* of informativeness can arise through iterated learning as a side-effect of applying a theoretically well-motivated simplicity principle during learning. Indeed, Richie (2016, p. 457) recently made a theoretical argument along these lines in which he describes the emergence of informativeness from the “basic operation of categorization” as a “happy accident.”

In Carstensen et al.’s (2015) studies, the randomly generated category systems that initialize the chains are already maximally expressive and maximally well-balanced, but they are not maximally compact. Therefore, the only way communicative cost can decrease in those experiments is through a generational increase in compactness, which may be explained in three ways. First, participants could discover through interaction that compact structures minimize error, and therefore switch to using such systems. This has been demonstrated by Jäger and van Rooij (2007), for example, who showed in computer simulations that convex (compact) color concepts can arise from a desire to minimize potential error during interaction. However, this explanation can be ruled out in the case of Carstensen et al. (2015) because neither of their studies had an interactional component. Second, participants could have a bias for informativeness in learning—as Carstensen et al.

## SIMPLICITY AND INFORMATIVENESS IN SEMANTIC CATEGORY SYSTEMS 30

(2015) propose—that is sensitive to the compactness property; the effects of this bias are then amplified by iterated learning. Or third, participants could have some other bias—we have proposed a domain-general simplicity bias—that also happens to favor compactness.

As we saw in the model (see Fig. 7), communicative cost is indeed expected to decrease under a simplicity bias, at least initially, because the simplicity prior favors contiguous categories which register as being more compact and therefore less costly; but then the loss of expressivity comes into effect and communicative cost begins to rise again. In our iterated learning experiment, the increase in contiguity combined with the loss of expressivity appear to cancel each other out, resulting in flat results for communicative cost (see Fig. 16). This raises the question as to why Carstensen et al. (2015) do not observe category loss, the answer to which appears to be different in each of their two studies.

The experimental design of Study 1 (i.e., J. Xu et al., 2013) explicitly forces participants to use a certain number of categories according to condition, so there can be no category loss over generations. In Study 2, there is no bottleneck on transmission—participants are trained on all 71 meanings; as such, category loss is unlikely, especially over the course of just 10 generations. To illustrate this, we reran our model approximating the parameters of Carstensen et al.’s (2015) Study 2: Agents have an unweighted ( $w = 1$ ) simplicity prior and see all of the meanings ( $b = 4$ , no bottleneck) in two exposures ( $\xi = 2$ ). Fig. 19 plots results for expressivity and communicative cost under three noise values. If we look only at the first 10 generations, there appears to be a small but sustained decrease in communicative cost, falling from 5 bits to around 4.7 bits,<sup>10</sup> which suggests that the languages are becoming more informative. However, over longer periods of time category loss begins to set in and the gain in informativeness is gradually eroded. This is especially clear when the noise parameter is raised (e.g.,  $\epsilon = .1$ ), which causes this process to happen faster.

When category loss is impeded, the only way languages can simplify is by making categories more compact (grouping stimuli together based on similarity), which in turn makes the languages appear more informative. Crucially, we argue that this outcome—increased informativeness—is *not* due to an inductive bias to improve communicative accuracy; it arises merely as a byproduct of increased simplicity. Or, at the very least, we argue that this is the more parsimonious explanation. It could be argued that some mixture of the two priors might offer a better fit to the experimental data—perhaps humans have biases for both simplicity and informativeness. However, the point we are making in this paper is that a simplicity bias alone can already explain the decrease in communicative cost observed by Carstensen et al. (2015), and since a domain-general bias for simplicity is very well motivated theoretically, this explanation should be preferred over positing an additional domain-specific bias for informativeness in language learning. This effectively means we are raising the bar on the evidence required to show that learning produces informative languages; evidence for this position must first rule out explanations from simplicity.

In this paper we treat learning independently of interaction, whereas learning often takes place in the context of interacting—in the context of trying to accomplish some goal. By treating the two separately, we can elucidate the unique contribution that each

<sup>10</sup>Compare with Carstensen et al. (2015, Fig. 5), where the decrease in communicative cost is similarly small, dropping from around 5.8 bits to 5.5 bits. Their numbers are slightly different from ours because they have more meanings (71 vs. 64) and the similarity between meanings is calculated differently.

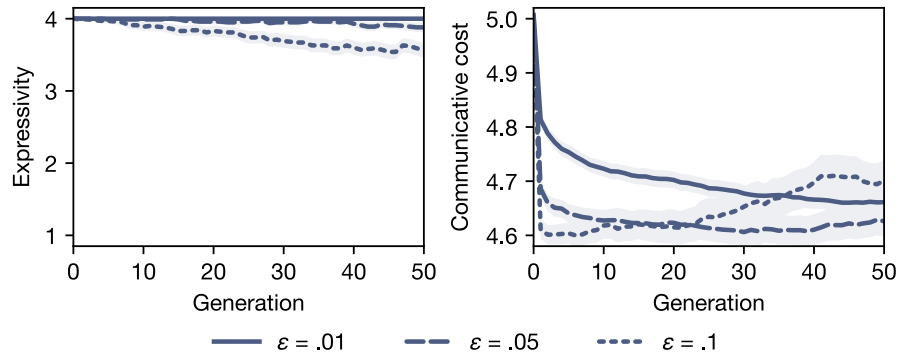


Figure 19. Model results under the simplicity prior for three settings of  $\epsilon$ . As in Carstensen et al. (2015, Study 2), agents see all of the meanings ( $b = 4$ ) in two exposures ( $\xi = 2$ ). When there is no bottleneck on transmission, category loss occurs slowly, making it appear that iterated learning gives rise to informativeness over 10 generations. However, category loss—and therefore the erosion of the informativeness gain—is inevitable if  $\epsilon > 0$ . Increasing the noise level speeds this process up.

pressure makes to the structure of language, this paper being primarily concerned with the contribution that learning makes. That being said, considering the two pressures together reveals the interesting way in which learning and interaction affect each other. Frank, Goodman, Lai, and Tenenbaum (2009, p. 1228) argue that “communicators choose what they want to say by how informative it would be about their intended meaning;” thus, the data from which learners typically induce simple hypotheses is often explicitly designed, by the speaker, to be informative in a given context. This suggests an important role for pragmatics in shaping language and that the true source of informativeness may lie in the production of data for an audience (as argued by Kirby et al., 2015) and in a given context (as argued by Winters et al., 2018). In this sense, informativeness does derive from a cognitive source, but that source is pragmatic reasoning, not learning.

### Conclusion

In a variety of studies, Regier and colleagues have found that communication systems find an optimal balance between simplicity and informativeness. What are the principal forces that lead to this? One possible explanation is that induction favors simplicity, while interaction favors informativeness; when languages are learned and used in this context, they become optimized under the simplicity–informativeness tradeoff. Alternative accounts place greater emphasis on the biases active during learning, suggesting that informativeness arises from a learning principle rather than from the dynamics present in interaction.

The findings we present in this paper challenge this alternative account. We do indeed find that semantic categories can become slightly more informative through iterated learning, but only in the limited sense that there is a general increase in compactness, due to a bias for simplicity, which in turn registers as an increase in informativeness. Our model



## SIMPLICITY AND INFORMATIVENESS IN SEMANTIC CATEGORY SYSTEMS 32

shows that a cognitive bias for informativeness only leads to learnable (i.e., transmissible) category structures under a strengthened form of Regier and colleagues' communicative cost measure which exaggerates the compactness component. But, even then, it results in a prediction that learners will find two-dimensional categories natural to learn (not supported by Experiment 1), which will be amplified in iterated learning (not supported by Experiment 2). In contrast, our model predicts that a simplicity bias will lead to category loss and contiguous categories that mark distinctions on only one dimension, findings that are strongly supported by our second experiment.

We maintain that language is best understood as arising from the tradeoff between simplicity and informativeness. However, when it comes to the compactness property of semantic category systems, the tradeoff does not apply; compact categories are beneficial to both learning and use, making it difficult to identify and test causal explanations. The argument we have put forward in this paper rules out one potential explanation—that compactness derives from a learning bias for informativeness—but two other potential explanations remain: Are categories compact because of an inductive bias for simplicity or because of the dynamics involved in true interaction?

### Acknowledgments

JWC was funded by the Economic and Social Research Council (grant number ES/J500136/1).

### References

- Ashby, F. G., & Maddox, W. T. (1990). Integrating information from separable psychological dimensions. *Journal of Experimental Psychology: Human Perception and Performance*, 16, 598–612. doi: 10.1037/0096-1523.16.3.598
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67, 1–48. doi: 10.18637/jss.v067.i01
- Beckner, C., Pierrehumbert, J. B., & Hay, J. (2017). The emergence of linguistic structure in an online iterated learning task. *Journal of Language Evolution*, 2, 160–176. doi: 10.1093/jole/lzx001
- Canini, K. R., Griffiths, T. L., Vanpaemel, W., & Kalish, M. L. (2014). Revealing human inductive biases for category learning by simulating cultural transmission. *Psychonomic Bulletin & Review*, 21, 785–793. doi: 10.3758/s13423-013-0556-3
- 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, 892–923. doi: 10.1111/cogs.12371
- Carstensen, A., Xu, J., Smith, C. T., & Regier, T. (2015). Language evolution in the lab tends toward informative communication. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 303–308). Austin, TX: Cognitive Science Society.
- Chater, N., Clark, A., Goldsmith, J. A., & Perfors, A. (2015). *Empiricism and language learnability*. Oxford, UK: Oxford University Press. doi: 10.1093/acprof:oso/9780198734260.001.0001

## SIMPLICITY AND INFORMATIVENESS IN SEMANTIC CATEGORY SYSTEMS 33

- Chater, N., & Vitányi, P. M. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences*, 7, 19–22. doi: 10.1016/S1364-6613(02)00005-0
- Cheung, H.-n. S. (1990). Terms of address in Cantonese. *Journal of Chinese Linguistics*, 18, 1–43.
- Culbertson, J., & Kirby, S. (2016). Simplicity and specificity in language: Domain-general biases have domain-specific effects. *Frontiers in Psychology*, 6, 1–11. doi: 10.3389/fpsyg.2015.01964
- Eppstein, D. (2010). Graph-theoretic solutions to computational geometry problems. In C. Paul & M. Habib (Eds.), *Graph-theoretic concepts in computer science* (pp. 1–16). Berlin, Germany: Springer. doi: 10.1007/978-3-642-11409-0\_1
- Fass, D., & Feldman, J. (2002). Categorization under complexity: A unified MDL account of human learning of regular and irregular categories. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems 15* (pp. 35–42). Cambridge, MA: MIT Press.
- Fedzechkina, M., Jaeger, T. F., & Newport, E. L. (2012). Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences of the USA*, 109, 17897–17902. doi: 10.1073/pnas.1215776109
- Feldman, J. (2016). The simplicity principle in perception and cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 7, 330–340. doi: 10.1002/wcs.1406
- Frank, M. C., Goodman, N., Lai, P., & Tenenbaum, J. B. (2009). Informative communication in word production and word learning. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 1228–1233). Austin, TX: Cognitive Science Society.
- Gärdenfors, P. (2000). *Conceptual spaces: The geometry of thought*. Cambridge, MA: MIT Press.
- Gärdenfors, P. (2014). *The geometry of meaning: Semantics based on conceptual spaces*. Cambridge, MA: MIT Press.
- Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science*, 31, 441–480. doi: 10.1080/15326900701326576
- Head, T., MechCoder, Louppe, G., Shcherbatyi, I., fcharras, Vinícius, Z., ... Fabisch, A. (2018). *scikit-optimize: v0.5.1*. doi: 10.5281/zenodo.1170575
- Jäger, G., & van Rooij, R. (2007). Language structure: Psychological and social constraints. *Synthese*, 159, 99–130. doi: 10.1007/s11229-006-9073-5
- Kemp, C. (2012). Exploring the conceptual universe. *Psychological Review*, 119, 685–722. doi: 10.1037/a0029347
- Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, 336, 1049–1054. doi: 10.1126/science.1218811
- Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics*, 4, 109–128. doi: 10.1146/annurev-linguistics-011817-045406
- Khetarpal, N., Neveu, G., Majid, A., Michael, L., & Regier, T. (2013). Spatial terms across languages support near-optimal communication: Evidence from Peruvian Amazonia, and computational analyses. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 764–769). Austin, TX: Cognitive Science Society.

## SIMPLICITY AND INFORMATIVENESS IN SEMANTIC CATEGORY SYSTEMS 34

- 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 of the USA*, *105*, 10681–10686. doi: 10.1073/pnas.0707835105
- Kirby, S., Griffiths, T. L., & Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, *28*, 108–114. doi: 10.1016/j.conb.2014.07.014
- Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102. doi: 10.1016/j.cognition.2015.03.016
- Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago, IL: The University of Chicago Press.
- Levinson, S. C. (2012). Kinship and human thought. *Science*, *336*, 988–989. doi: 10.1126/science.1222691
- Li, M., & Vitányi, P. M. (2008). *An introduction to Kolmogorov complexity and its applications*. New York, NY: Springer. doi: 10.1007/978-0-387-49820-1
- Martinet, A. (1952). Function, structure, and sound change. *Word*, *8*, 1–32. doi: 10.1080/00437956.1952.11659416
- Meilă, M. (2007). Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, *98*, 873–895. doi: 10.1016/j.jmva.2006.11.013
- Moreton, E., Pater, J., & Pertsova, K. (2015). Phonological concept learning. *Cognitive Science*, *41*, 4–69. doi: 10.1111/cogs.12319
- Motamedi, Y., Schouwstra, M., Culbertson, J., Smith, K., & Kirby, S. (in press). Evolving artificial sign languages in the lab: From improvised gesture to systematic sign. *Cognition*.
- Murdock, G. P. (1970). Kin term patterns and their distribution. *Ethnology*, *9*, 165–207. doi: 10.2307/3772782
- Murphy, G. L. (2004). *The big book of concepts*. Cambridge, MA: MIT Press.
- Nielsen, A., & Rendall, D. (2012). The source and magnitude of sound-symbolic biases in processing artificial word material and their implications for language learning and transmission. *Language and Cognition*, *4*, 115–125. doi: 10.1515/langcog-2012-0007
- Nygaard, L. C., Cook, A. E., & Namy, L. L. (2009). Sound to meaning correspondences facilitate word learning. *Cognition*, *112*, 181–186. doi: 10.1016/j.cognition.2009.04.001
- Raviv, L., Meyer, A., & Lev-Ari, S. (2018). The role of community size in the emergence of linguistic structure. In C. Cuskley, M. Flaherty, H. Little, L. McCrohon, A. Ravignani, & T. Verhoef (Eds.), *The evolution of language: Proceedings of the 12th international conference* (pp. 402–404). Toruń, Poland: Nicolaus Copernicus University. doi: 10.12775/3991-1.096
- Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences of the USA*, *104*, 1436–1441. doi: 10.1073/pnas.0610341104
- Regier, T., Kemp, C., & Kay, P. (2015). Word meanings across languages support efficient communication. In B. MacWhinney & W. O’Grady (Eds.), *The handbook of language emergence* (pp. 237–263). Hoboken, NJ: John Wiley & Sons. doi: 10.1002/9781118346136.ch11

## SIMPLICITY AND INFORMATIVENESS IN SEMANTIC CATEGORY SYSTEMS 35

- Richie, R. (2016). Functionalism in the lexicon: Where is it, and how did it get there? *The Mental Lexicon*, 11, 429–466. doi: 10.1075/ml.11.3.05ric
- Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14, 465–471. doi: 10.1016/0005-1098(78)90005-5
- Rosch, E. H. (1973). Natural categories. *Cognitive Psychology*, 4, 328–350. doi: 10.1016/0010-0285(73)90017-0
- Rosch, E. H. (1978). Principles of categorization. In E. H. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–48). Hillsdale, NJ: Lawrence Erlbaum.
- Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, 1, 54–87. doi: 10.1016/0022-2496(64)90017-3
- Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, 237, 1317–1323. doi: 10.1126/science.3629243
- Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, 75, 1–42. doi: 10.1037/h0093825
- Solomonoff, R. J. (1964). A formal theory of inductive inference. Part I. *Information and Control*, 7, 1–22. doi: 10.1016/S0019-9958(64)90223-2
- Spike, M., Stadler, K., Kirby, S., & Smith, K. (2017). Minimal requirements for the emergence of learned signaling. *Cognitive Science*, 41, 623–658. doi: 10.1111/cogs.12351
- Tamariz, M. (2017). Experimental studies on the cultural evolution of language. *Annual Review of Linguistics*, 3, 389–407. doi: 10.1146/annurev-linguistics-011516-033807
- von der Gabelentz, G. (1891). *Die Sprachwissenschaft: Ihre Aufgaben, Methoden und bisherigen Ergebnisse [Linguistics: Aims, methods, and current results]*. Leipzig, Germany: T.O. Weigel Nachfolger.
- Winter, B., & Wieling, M. (2016). How to analyze linguistic change using mixed models, Growth Curve Analysis and Generalized Additive Modeling. *Journal of Language Evolution*, 1, 7–18. doi: 10.1093/jole/lzv003
- Winters, J., Kirby, S., & Smith, K. (2018). Contextual predictability shapes signal autonomy. *Cognition*, 176, 15–30. doi: 10.1016/j.cognition.2018.03.002
- Xu, J., Dowman, M., & Griffiths, T. L. (2013). Cultural transmission results in convergence towards colour term universals. *Proceedings of the Royal Society B: Biological Sciences*, 280, 1–8. doi: 10.1098/rspb.2012.3073
- Xu, Y., & Regier, T. (2014). Numeral systems across languages support efficient communication: From approximate numerosity to recursion. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 1802–1807). Austin, TX: Cognitive Science Society.
- Xu, Y., Regier, T., & Malt, B. C. (2016). Historical semantic chaining and efficient communication: The case of container names. *Cognitive Science*, 40, 2081–2094. doi: 10.1111/cogs.12312
- Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge, MA: Addison-Wesley.

### 3.2 Summary of Paper 2

First and foremost, Paper 2 should be viewed as an attempt to understand – what we perceived as – a surprising result obtained by Carstensen et al. (2015). To reiterate, Carstensen et al. (2015) found that informative category systems can arise through iterated learning, despite the lack of any explicit communicative pressure. Although somewhat unclear, Carstensen et al.’s (2015) position appears to be that learners expect languages to be informative and are therefore equipped with a bias for informative languages; the effects of this bias are then amplified by the process of iterated learning. In particular, Carstensen et al. (2015) make reference to Fedzechkina et al. (2012), describing that work as establishing ‘the general principle that learners may alter their input in the direction of greater efficiency’, where by ‘efficiency’ they appear to mean in terms of communicative interaction.

Consequently, the approach we took in the paper was to formalize two possible assumptions about the bias that learners bring to the table – a bias for simplicity or a bias for informativeness – and then test the two formalizations experimentally. Doing so revealed that a bias for simplicity offers a very strong explanation for not only *our* experimental results but also those of Carstensen et al. (2015), as highlighted in Fig. 19 on page 82. Specifically, Carstensen et al.’s (2015) findings can *only* be explained by an increase in compactness,<sup>11</sup> which is a hallmark feature of both simplicity and informativeness; we argue, therefore, that the authors have attributed to informativeness something that is more parsimoniously attributed to simplicity given that their studies only include a pressure from learning. The upshot of this is that iterated learning can indeed give rise to more informative category systems, but only in the rather limited sense that the categories become more compact, which is not due to optimization for communicative efficiency but due to a preference for simplicity in induction.<sup>12</sup>

Indeed, Richie (2016, p. 457) – whose article I only found during the writeup of Paper 2 – appears to have come to the same conclusion. He says, ‘Carstensen et al. have found that iterated learning transforms randomly partitioned color and spatial re-

11 Their initial, randomly-generated languages start out maximally expressive and maximally well-balanced, so the only way communicative cost can decrease over time, as they observed, is through an increase in the compactness of the categories.

12 Which itself can be motivated by cognitive economy (‘compressed representations’) or the application of Occam’s razor (‘compressible explanations’). See Section 2.2.3.

lation lexicons into lexicons of greater informativeness', and after reviewing some of Gärdenfors's ideas on convexity, he goes on to say, 'it may be that this particular functional aspect of lexicons follows not from pressures for utilitarian communication or categorization, but merely from the basic operation of categorization itself, making this aspect of lexicons a "happy accident"'. In other words, the fact that compact categories are good for functional communicative reasons may in fact be an accident, with cognitive principles of categorization and learning being the real source behind such compact structure. If correct, this casts a different light on, for example, Jäger and van Rooij's (2007) computer simulations, which showed that convex category structures emerge from interaction.

One criticism that could be levelled at the paper is that the stimuli used in our experiments are quite different from those used in Carstensen et al. (2015), especially in terms of the integral–separable distinction. 'Integral dimensions are those that combine into relatively unanalyzable, integral wholes' (Nosofsky, 1986, p. 40); for example, when perceiving colour, humans integrate information about hue, saturation, and brightness simultaneously rather than perceive the three dimensions separately. In contrast, 'separable dimensions are highly analyzable and remain psychologically distinct when in combination' (Nosofsky, 1986, p. 40), the Shepard circles being classic examples of such stimuli (the angle and size components can be perceived and categorized separately). Both of Carstensen et al.'s (2015) studies use stimuli with integral dimensions (colour and spatial relationships), which raises the possibility that our experimental results might not offer a fair comparison to their work. In particular, one might argue that, if we were to adopt stimuli with integral dimensions, participants in Experiment 1 would find the quadrant partition just-as-easy or easier to learn than the striped partition, and, as such, the iterated learning chains in Experiment 2 would converge on more quadrant-like partitions.

My response to this criticism is as follows. Our model of the informativeness prior is actually closer to the integral case because we adopt the Euclidean metric, which is argued to be more representative of integral dimensions than separable dimensions (Nosofsky, 1986; Shepard, 1964).<sup>13</sup> Therefore, our model of the informativeness prior would remain unchanged in the face of this argument and the results would therefore

<sup>13</sup> Incidentally, switching to the Manhattan metric – argued to be a better model of separable dimensions – yields the same overall results; the predictions of the model are indifferent to the choice of distance metric.

be the same: The informativeness prior favours the quadrant partition (because of its compact packing), so long as the compactness component is weighted strongly enough. Our model of the simplicity prior would have to be revised, however, since the complexity measure based on the rectangle code is suited to modelling separable dimensions. It is unclear what this revised complexity measure might look like precisely, but, for the sake of argument, let's say it would also favour the quadrant partition (perhaps, for example, this revised measure would be based on representing concepts as prototypes, ultimately favouring compact packings). This would then leave us with a problem. Both priors would make the same prediction regarding the geometric structure of the concepts, making it difficult – or even impossible – to determine which was the better model via Experiment 1. Nevertheless, this revised simplicity prior would still favour having fewer categories than many, so the model would still predict a loss of expressivity over generational time, which ultimately means that languages are predicted to become *less* informative through iterated learning, not *more* informative as suggested by Carstensen et al. (2015). In short, our choice of separable stimuli was a deliberate one to make it easier to distinguish between the two theoretical predictions.

To address this concern even more directly, even if we were to rerun our iterated learning experiment with exactly the same stimuli used in Carstensen et al.'s (2015) Study 2, our conclusions would remain the same. We would expect to see a brief initial increase in informativeness (as the concepts reorganize into simpler compact arrangements), followed by a dramatic loss of informativeness as conceptual distinctions are lost in favour of simplicity. To reiterate, the reason why this does not happen in Carstensen et al. (2015), at least not within the first ten generations, is because they do not apply a bottleneck on intergenerational transmission, which has the effect of slowing down the loss of expressivity and making it appear as though iterated learning gives rise to informativeness.

In the remainder of this chapter, I discuss other aspects of the project that were not discussed in the paper. First, in Section 3.3, I describe work attempting to simulate the process of communication between participants who took part in Experiment 1. In Sections 3.4 and 3.5, I provide more detail on technical aspects of the model: the implementation of the rectangle code and two proposal functions used to sample from the hypothesis space. Section 3.6 provides a proof of concept of the model-fit procedure,

demonstrating that model parameters can successfully be recovered from simulated iterated learning results. And, finally, Section 3.7 concludes the chapter.

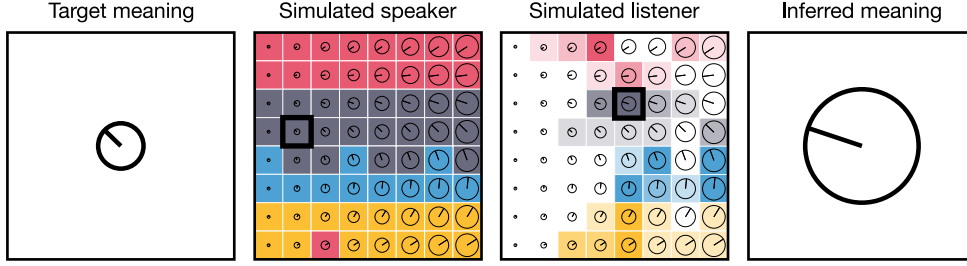
### 3.3 Simulating Participant Communication

In Experiment 1, we collected production and comprehension responses from participants. Although not mentioned in the paper, one of the reasons we collected both types of response was in order to simulate communication. In such a simulation, a participant from the Production test type becomes a speaker who produces signals given meanings, and a participant from the Comprehension test type becomes a listener who infers meanings given signals. This method offers a number of advantages over running a true communication experiment. Most importantly it means that we can isolate the two facets of communication to understand what happens on each side without the interference that occurs when participants are continually updating their behaviour in response to each other.

To perform these simulations, a participant from the Production version of Experiment 1 (the simulated speaker) is randomly paired with a participant from the Comprehension version (the simulated listener); both participants had been taught the same system (Angle-only, Size-only, or Angle & Size). An interaction then takes place in which the simulated speaker expresses a signal for a meaning chosen at random, and the simulated listener infers some meaning in response. More specifically, given the randomly selected meaning thrown up by the world, the simulated speaker utters a signal according to how that meaning was labelled by the participant. On reception of this signal, the simulated listener responds by sampling a meaning from the meanings that the participant choose as examples of that signal. The success of the interaction is then measured as the Euclidean distance between the target and inferred meanings. An example is shown in Fig. 3.1; see also Appendix C for individual participant languages.

The density plots on the left-hand side of Fig. 3.2 show results from 10,000 simulated interactions under each of the three category systems. These results do *not* support the idea that the two-dimensional, Angle & Size partition minimizes communicative error; under this system, communicative error tended to be higher. However, the reason for this is that participants in that condition tended to learn the system very poorly in the first place. In other words, any advantage that the Angle & Size system had in terms of





**Figure 3.1:** Illustration of a simulated interaction. The world throws up some target meaning, and the simulated speaker (i.e. a participant who was assigned to the Production test type) utters the signal associated with that meaning – in this case, the signal associated with the grey category. On reception of this signal, the simulated listener (i.e. a participant who was assigned to the Comprehension test type) infers a meaning; the simulated listener is more likely to choose a meaning that the participant repeatedly selected as an example of the grey category (i.e. meanings in a darker shade of grey). In this case, the outcome of the interaction is not fully successful; the communicative error (Euclidean distance between target and inferred meanings) is 3.16.

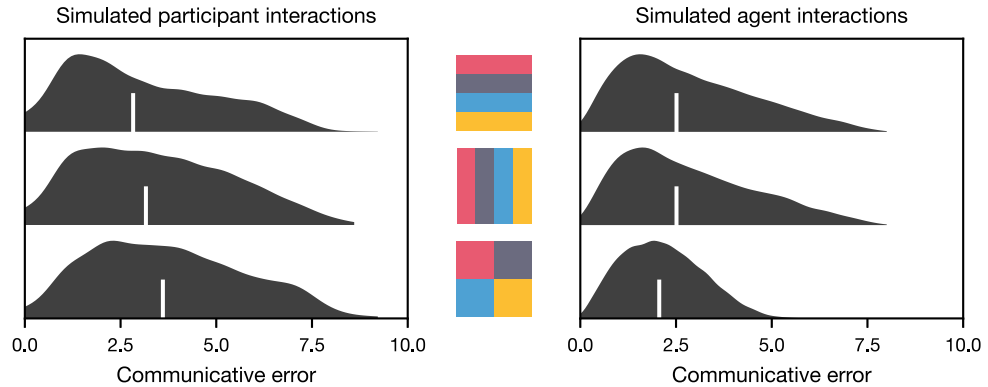
communicative accuracy is overpowered by the difficulty involved in learning it. Nevertheless, we can still simulate the level of communicative accuracy that is expected under each system given communicators who have a perfect grasp of the system. These results are shown on the right-hand side of Fig. 3.2 and are based on the assumption that listeners have a Gaussian representation of the categories – that they are more likely to select category central meanings.<sup>14</sup> As predicted, communicative error is minimized under the Angle & Size system since the categories have a more compact packing.

### 3.4 Complexity and the Rectangle Code

In order to implement the simplicity bias, Paper 2 adopts part of the method proposed by Fass and Feldman (2002) who made use of the MDL principle to formulate a model of concept induction (see Section 2.2.5). As we saw in Section 2.2.4, the MDL principle is mathematically equivalent to Bayes' theorem. As such, it is possible to mix an MDL-based prior with a standard likelihood function. Our model with the simplicity bias does exactly this; the simplicity prior is defined on page 57 as

$$\pi_{\text{sim}} \propto 2^{-\text{complexity}(L)}, \quad (3.1)$$

<sup>14</sup> However, even if we do not make this assumption, the same basic result holds because the quadrant partition is more compact than the striped partitions.



**Figure 3.2:** Density plots showing communicative error from 10,000 simulated interactions between participants (left) and ideal agents (right). Error is measured as the Euclidean distance between intended and inferred meanings. The results from the actual participants (left) do not support the idea that the two-dimensional, Angle & Size partition minimizes communicative error; however, this is in fact because participants in that condition tended to learn the system very poorly in the first place. The equivalent results from the ideal agents (right) shows that communicative error is indeed minimized under the Angle & Size system.

which simply transforms the binary description length (complexity) into a probability by negative exponentiation. The likelihood of the data given a hypothesis is then calculated in a more standard probabilistic way (see page 57), rather than formulating an MDL-based likelihood (as is the case in Fass & Feldman, 2002). The reason we did not formulate the model entirely in MDL terms is because we want to compare two prior functions (the informativeness prior is not formulated in MDL terms), while keeping the likelihood function constant. Alternatively, our approach to the simplicity prior can be thought of as an approximation of Solomonoff’s universal prior – the prior probability of a language is inversely proportional to the complexity of that language, and we use the rectangle code as a convenient measure of complexity.

### 3.4.1 Deviations from Fass & Feldman’s rectangle code

The rectangle code is used to describe a category’s extension in terms of a set of rectangles that minimize description length (see pages 57–60). The method we use differs from Fass and Feldman (2002) in a few small ways.

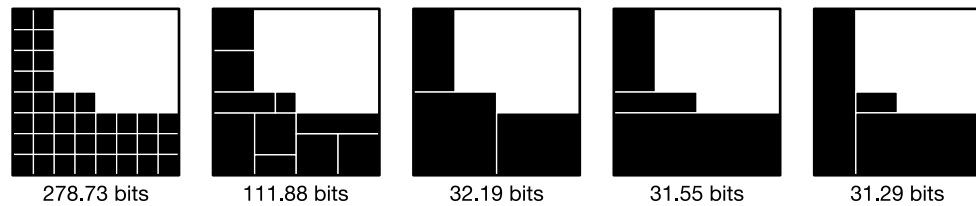
Firstly, our model and experiment allow languages to consist of up to four categories, while in Fass and Feldman (2002) there are two categories that must be learned (ally vs. enemy) or arguably just one category that must be learned – enemy ships, the

complement of which constitutes the ally ships (or *vice versa*). Thus, in Fass and Feldman (2002), complexity is measured on a single category, while in our case, we sum the complexities of up to four categories to arrive at a measure of complexity for the language as a whole. Technically, this simple summation approach is not correct. When we formulate a description of a *category*, we concatenate the binary codewords of each rectangle symbol that is needed to describe that category; since the rectangle code is prefix-free, this description may be unambiguously decoded. However, when we formulate a description of a *language*, the category descriptions are concatenated in a way that cannot be decoded, since there is no separator symbol to mark where one category ends and the next begins. In Fig. 2.6 (page 33), for example, we gloss over this issue by highlighting the different category descriptions in different colours. We did consider devising a description scheme for entire languages – either by introducing an additional separator symbol or by introducing four new symbols, one to mark the start of each category – but we felt that doing so would needlessly complicate the method with little benefit beyond the simple summation/concatenation approach. Such a method would essentially add a fixed number of bits for each additional category, further penalizing languages that have more categories.

Secondly, Fass and Feldman (2002) adopt a fully continuous space in their experiment, such that there is, in theory, an infinite number of stimuli that participants could be exposed to; this continuous space is then quantized onto a  $4 \times 4$  grid, and up to four rectangles are selected that are representative of the how the participant treated the category's extension. For us, the continuous space was first quantized onto an  $8 \times 8$  grid, such that participants are only ever exposed to 64 particular stimuli. This was designed to make certain other things easier, such as bottlenecking and multiple exposures (although Paper 3 in the next chapter offers a method for using a fully continuous space in iterated learning experiments). It also means that we can directly find the exact set of rectangles required to represent a category's extension.

### 3.4.2 Rectangular decomposition

Estimating the complexity of a category using the rectangle code is nontrivial. In particular, finding the *shortest* possible description of a category is a difficult search problem, especially as the size of the category grows. These issues were only very briefly alluded

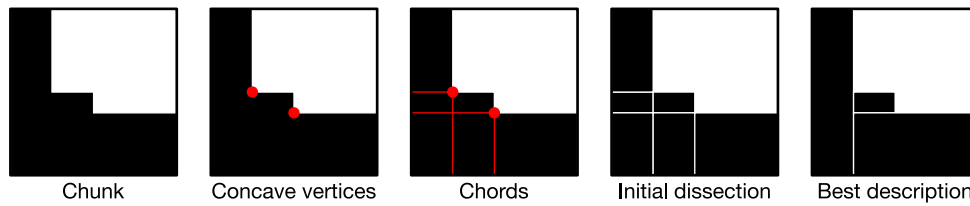


**Figure 3.3:** Five ways in which a category, modelled as a rectilinear polygon (i.e. the black area), may be described in terms of a set of rectangles. In the first case, the category is described in terms of 36  $1 \times 1$  rectangles, yielding a description length of 278.73 bits. The second case takes advantage of the fact that fewer than 36 rectangles may be used to describe the category in a more compressed way. For this particular category, a minimum of three rectangles is required, but multiple descriptions are still possible, some shorter than others.

to in Footnote 3 (page 57), so in this section I described these methods more fully.

Fig. 3.3 illustrates five ways in which an example category – essentially a rectilinear polygon – could be decomposed into a set of rectangles. Finding the minimal number of rectangles needed to represent a polygon has practical applications in, for example, microprocessor design and the compression of bitmap images, and has been extensively studied in the field of computational geometry. Lipski Jr, Lodi, Luccio, Mugnai, and Pagli (1979), Ohtsuki (1982), and Ferrari, Sankar, and Sklansky (1984) independently discovered a graph-theoretic, polynomial-time algorithm to do exactly this (see Eppstein, 2010, for a brief introduction). Roughly, this algorithm proceeds in five main steps:

1. Identify the polygon's concave vertices.
2. Identify any vertical or horizontal chords that link two concave vertices.
3. Identify any chords that intersect, yielding a bipartite graph since only vertical and horizontal chords may intersect.
4. Select as many of the chords as possible that do not intersect. This is equivalent to finding a maximum independent set in the bipartite graph, which may be solved by the Hopcroft–Karp algorithm (Hopcroft & Karp, 1973).
5. For any remaining concave vertices, make an arbitrary decision about which of two slices to make.



**Figure 3.4:** Illustration of the decomposition process. A category may comprise multiple chunks (contiguous rectilinear polygons), each of which is treated independently. During decomposition, we identify the chunk's concave vertices and the two chords that emanate from each; we then dissect the polygon into an initial set of rectangles, which may then be merged to form the shortest possible description.

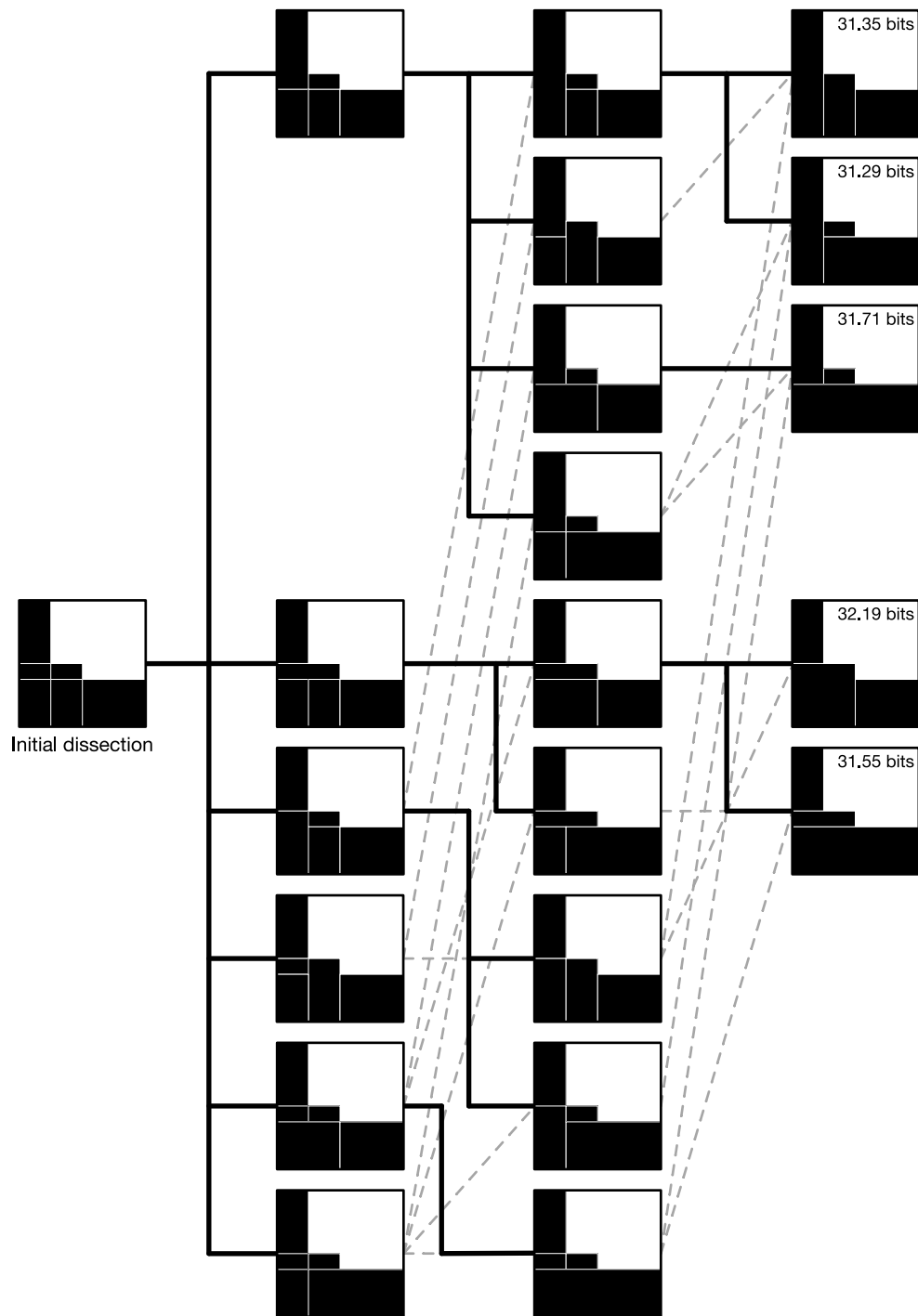
This algorithm yields a rectangularization that is comprised of the minimum number of rectangles, but it will not necessarily have the shortest possible description length. In Fig. 3.3, for example, the minimum number of rectangles is three, but the description lengths are different depending on which one is chosen.

The method used in the paper combines some of the techniques described above with a brute-force approach, and proceeds as follows (illustrated in Fig. 3.4):

1. Separate the concept into 'chunks' (contiguous rectilinear polygons), since each chunk can be treated as a separate decomposition problem.
2. For each chunk, identify its concave vertices.
3. Identify the two chords emanating from each concave vertex, and dissect the chunk into an initial set of rectangles.
4. Merge rectangles that share a complete edge until no further mergers are possible and find a rectangularization that minimizes description length.

Step 4 in this process uses a simple recursive function of the form:

```
Function Rectangularize(rectangle_set):
  For each pair of rectangles in rectangle_set:
    If the pair share an edge:
      Merge the pair of rectangles into one rectangle
      Calculate the description length of new_rectangle_set
      If shorter than the shortest observed so far:
        Store new_rectangle_set and its description length
      Pass new_rectangle_set into the Rectangularize function
```



**Figure 3.5:** Beginning with the initial dissection (left), the recursive merge algorithm explores each sequence of merging pairs of rectangles. In this case, the algorithm goes up to three levels of recursion deep, yielding five rectangularizations that are comprised of the minimum number of rectangles. Of these, it selects a rectangularization that minimizes description length. When the algorithm hits a previously-evaluated candidate (indicated by dashed grey lines), it may safely ignore the candidate and its descendants, greatly reducing the number of candidates that must be evaluated.

This recursive function exhaustively explores all possible sequences in which rectangles in the initial dissection could be merged together, as illustrated in Fig. 3.5. This is necessary, since the order in which rectangles are merged affects the ultimate outcome. However, a memoization technique is also applied to avoid evaluating the same candidate solution multiple times, which has the effect of eliminating entire branches of the search tree and greatly reducing the set of candidate rectangularizations. Even so, the number of candidate solutions that must be considered becomes impractical with a initial dissection of more than 20 rectangles and intractable with a initial dissection of more than 30 rectangles. Although such cases are fairly rare in our data, where they occur we use a beam search variant on the above algorithm in which we only recurse on the most promising branches, which is not guaranteed to find the shortest description.

### 3.5 Metropolis–Hastings and the Proposal Function

To sample a language from the posterior distribution, our model uses the Metropolis–Hastings algorithm (see pages 61–62). This is because the set of possible languages is too large ( $4^{64}$ ) to calculate the posterior probability in each case. Our implementation of the Metropolis–Hastings algorithm proceeds in the following way:

1. Generate a random language (a random labelling of the  $8 \times 8$  space).
2. Calculate the posterior probability of that language.
3. Propose a new candidate language by mutating its current state.
4. Calculate the posterior probability again.
5. Calculate the acceptance ratio,  $\alpha$ ; if  $\alpha \geq 1$ , accept the candidate language automatically; if  $\alpha < 1$ , accept the candidate with probability  $\alpha$ .
6. Repeat steps 3–5 a large number of times, after which the final state is taken to be a representative draw from the true posterior distribution.

The reason Metropolis–Hastings yields a fair draw is because, in the limit, the number of iterations spent on a given language is proportional to that language’s posterior probability. Crucially for our purposes, Metropolis–Hastings is able to do this without

knowledge of the true posterior probability of a language, the calculation of which involves an intractable normalization term; in Metropolis–Hastings, it is only the ratio between posterior probabilities that matters, not their true values.

As reported in the paper (page 62), the acceptance ratio  $\alpha$  is given by

$$\alpha = \frac{p(L'|D)}{p(L_i|D)} \cdot \frac{p(L_i|L')}{p(L'|L_i)}, \quad (3.2)$$

where  $L_i$  is the current state of the language and  $L'$  is the proposed candidate. The acceptance ratio determines the probability that the algorithm will jump from the current state to another (the probability that it will accept the candidate), and it is defined as the product of two terms. The posterior ratio (on the left) tells us how many times more (or less) probable one language is compared to another in the (unnormalized) posterior distribution over languages. The proposal ratio (on the right) accounts for the baseline probability of proposing candidate  $L'$  given state  $L_i$ . Let's say the probability of proposing candidate  $L'$ ,  $p(L'|L_i)$ , is 0.1 and the probability of jumping back again,  $p(L_i|L')$ , is 0.09; the proposal ratio is  $0.09/0.1 = 0.9$ , which has the effect of slightly downweighting the posterior ratio, making the candidate language less likely to be accepted. This is desirable in this example case because the proposal function is slightly biased towards proposing state  $L'$ , which needs to be accounted for. Such proposal functions are described as asymmetric, since  $p(L_i|L') \neq p(L'|L_i)$ . In developing the model, I considered two possible proposal functions: *cell mutation*, which is symmetric, and *rectangle mutation*, which is asymmetric. In the paper, we use the asymmetric rectangle mutation method, but to understand why we will first look at cell mutation.

### 3.5.1 Cell mutation

In cell mutation, a new candidate language is proposed by selecting one of the 64 cells (meanings) at random in the  $8 \times 8$  space and changing its category membership to one of the other three possible categories (at random). This function is symmetric,  $p(L_i|L') = p(L'|L_i)$ , because the probability of proposing any particular candidate given some current state is always  $(1/64)(1/3)$ . Therefore, the proposal ratio is equal to 1 and drops out of Equation 3.2, simplifying to the Metropolis algorithm. Although this makes the calculations easier and is more transparent, cell mutation is problematic because it is prone to getting stuck in local maxima.





**Figure 3.6:** There exists some true language out in the world that an agent would like to infer. However, the agent is only able to observe a limited number of data points. Despite this impoverished dataset, an ideal learner may infer the true language by drawing on its prior knowledge of the types of language that tend to exist. In the optimal scenario, the agent would infer the true language; however, the agent could infer a suboptimal language, especially since there is little data to rely on in the bottom-left corner.

To illustrate this, consider the simple case shown in Fig. 3.6, where we use a  $4 \times 4$  space and allow for up to *two* categories, rather than four. The language that exists out in the world divides the space of meanings into two categories along the  $x$ -axis, and the agent observes data for half of the meanings. Despite this impoverished dataset, an ideal learner should be able to reconstruct the true language by weighing up the likelihood of the data under candidate languages and its prior knowledge of how probable such candidate languages are. However, given the situation illustrated in Fig. 3.6, runs of the Metropolis–Hastings algorithm (using cell mutation as the proposal function) often yield suboptimal solutions – languages that have high posterior probability, but not the highest across the entire hypothesis space – because they become stuck in local maxima.

An example of this is highlighted by Fig. 3.7, which depicts 80 iterations of the algorithm using cell mutation as the proposal function. After 36 iterations, the algorithm becomes stuck on a suboptimal language; the only way the algorithm can transition to the globally optimal state is by first transitioning to a lower probability state. This is because, under cell mutation, the algorithm is only able to change a single cell at a time. In the limit, the states considered by the Metropolis–Hastings algorithm will converge on the true posterior distribution; even if a candidate has lower posterior probability, there is still some small probability of it being accepted, allowing the algorithm to escape from local maxima; however, under cell mutation this can take an impractical amount of time, especially in the more complex case of an  $8 \times 8$  space with a greater number of possible categories. In other words, cell mutation is slow-mixing, requiring a very large



number of iterations to obtain a fair draw from the hypothesis space.

### 3.5.2 Rectangle mutation

My solution to the problem outlined above was to use a more complex mutation function that allows multiple cells to be changed at the same time. In rectangle mutation, a new candidate is proposed by selecting a rectangular region at random from the space (under the criterion that all meanings in that region must belong to a single category<sup>15</sup>) and changing its category membership to one of the other three possible categories (at random). The benefit of this mutation function is illustrated in Fig. 3.8; the algorithm is no longer prone to becoming stuck in local maxima because it is able to change multiple cells in a single step.

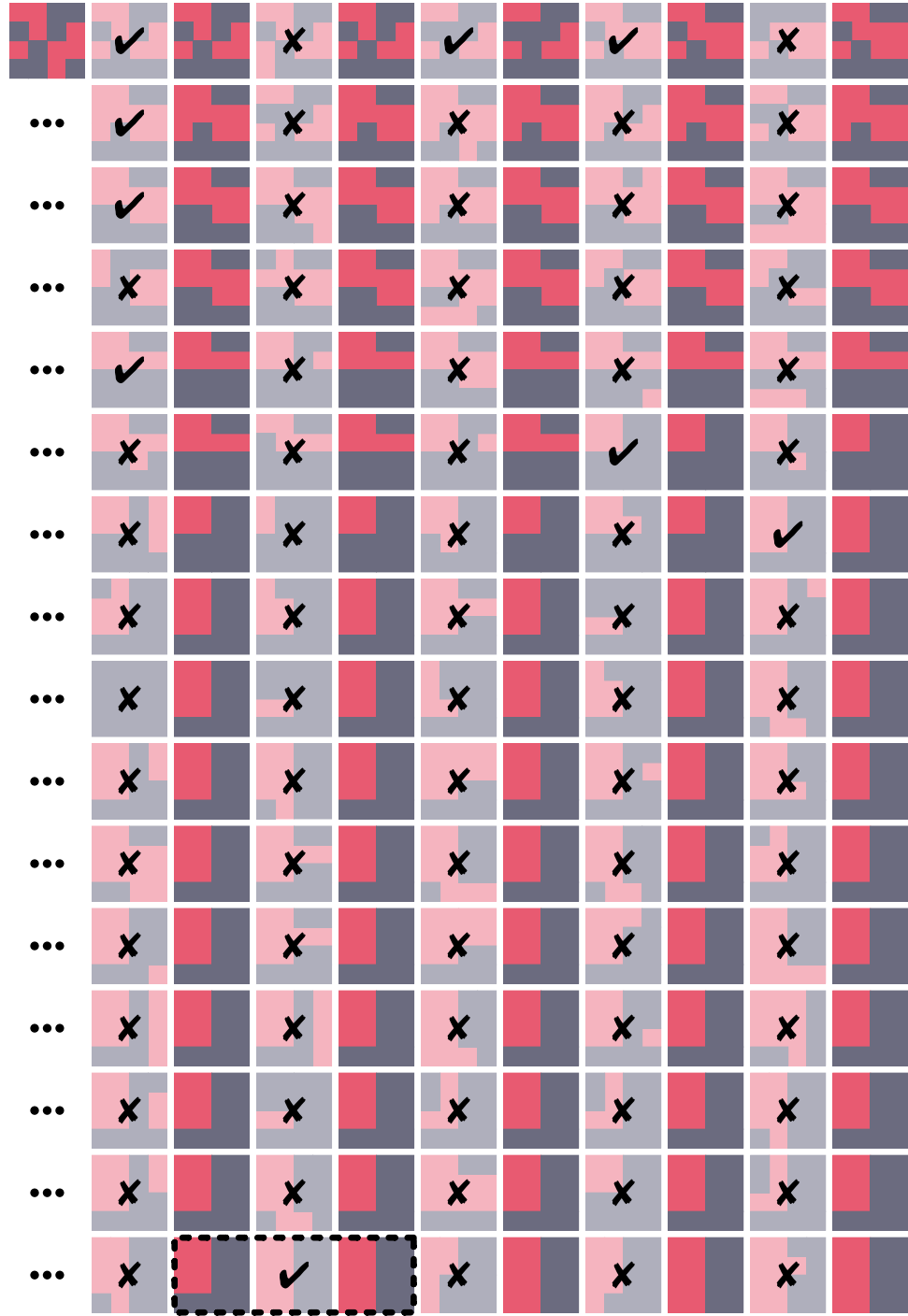
On each iteration of the algorithm, we enumerate a set of mutable rectangles – rectangles that do not cut across a category boundary and whose category membership may therefore be changed. Of  $n$  mutable rectangles, there is exactly one that allows the algorithm to jump from the current language  $L_i$  to the candidate language  $L'$ , so  $p(L'|L_i) = 1/n$ . Likewise, there is exactly one mutable rectangle that allows the algorithm to jump back again, although the number of mutable rectangles  $n'$  may be different. As such, rectangle mutation is asymmetric,  $p(L_i|L') \neq p(L'|L_i)$ , making it necessary to calculate the proposal ratio in Equation 3.2, which is simply  $(1/n')/(1/n)$ .

Finally, note that, although this proposal function uses rectangles, this process is not related to the rectangle code and decomposition methods described in the previous section; the use of rectangles in the proposal function is merely a convenient way to change multiple cells at each step in order to prevent the Metropolis–Hastings algorithm becoming stuck in local maxima. Indeed, the rectangle mutation method is used for both the simplicity prior and the informativeness prior.

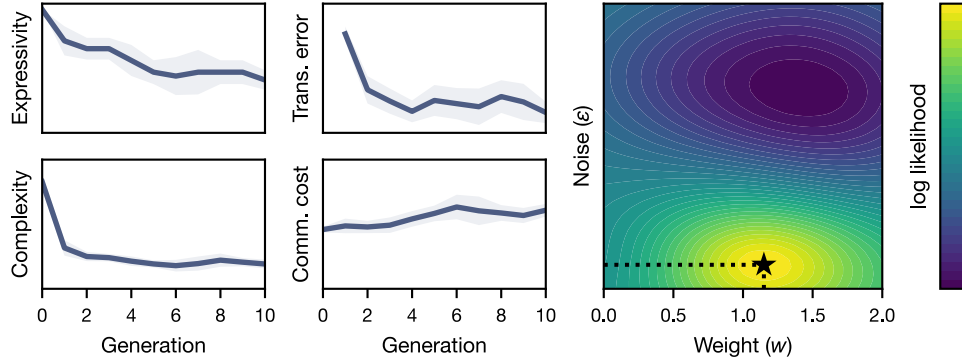
## 3.6 Testing the Model Fit Procedure

A crucial part of the paper is the section in which we fit the data from Experiment 2 to the model in order to determine which prior function offers the better fit (pages 76–

<sup>15</sup> Thus, the rectangle mutation method is not biased towards introducing new rectangles (artificially making the languages simpler) because it only modifies the category membership of rectangular regions that already existed at the previous step.



**Figure 3.8:** An MCMC chain under rectangle mutation. The first language is randomly generated and is followed by 80 iterations of the Metropolis–Hastings algorithm. Candidate languages are depicted in a lighter shade and are marked as accepted or rejected. In each of the candidates, all cells in a rectangular region that does not cut across a category boundary are transferred to the other category, and the candidate is accepted if this results in greater posterior probability. Under this mutation function, the algorithm can avoid becoming stuck in local maxima, since it is able to change the category membership of multiple cells in a single iteration (highlighted by the black box).



**Figure 3.9:** Simulated model fit used to test the model fit procedure. The iterated learning model was run using the simplicity prior with  $w$  set to 1.0 and  $\varepsilon$  set to 0.1. The results are shown on the left. The model fit procedure was then used to estimate the values of  $w$  and  $\varepsilon$  from the iterated learning results. The maximum likelihood estimates were  $w^* = 1.15$  and  $\varepsilon^* = 0.08$ , demonstrating that the model fit procedure is able to recover the model parameters from the resultant data.

78). This involved a fairly complex procedure that was highly computationally intensive, taking around one month to complete on a high-performance cluster. To evaluate the performance of the procedure, I applied it to simulated data to check that the model fit procedure was able to recover the weight and noise parameters from simulated results where the actual values of these parameters are known.

Specifically, a simplified version of the model was run using a  $4 \times 4$  space with ten chains of ten generations – effectively 100 virtual participants. The model parameters were set to:  $\pi = \pi_{\text{sim}}$ ,  $w = 1$ ,  $b = 2$ ,  $\xi = 2$ , and  $\varepsilon = 0.1$ . The results of this simulation are depicted on the left-hand side of Fig. 3.9; as expected under a simplicity prior, expressivity, transmission error, and complexity all decrease, while communicative cost begins to increase as categories are gradually lost. The model fit procedure was then used to estimate the values of  $w$  and  $\varepsilon$  from the 100 virtual participants'  $D_{\text{in}}-D_{\text{out}}$  pairs. In other words, for each of these virtual participants, we simulate what would happen when an agent learns from the virtual participant's  $D_{\text{in}}$  and then seek parameter values that maximize the likelihood of that agent producing the virtual participant's  $D_{\text{out}}$  (i.e. parameter values that maximize the probability of an agent producing the same output as the virtual participant when given the same input as that virtual participant).

The results are shown on the right-hand side of Fig. 3.9. The maximum likelihood estimates were  $w^* \approx 1.15$  and  $\varepsilon^* \approx 0.08$ . These values are very similar to the true parameter values ( $w = 1$  and  $\varepsilon = 0.1$ ), suggesting that the model fit procedure is indeed

able to recover the model parameters from the resultant data. That being said, the simulated results are much less noisy than the real world data, so it is unclear how well the model fit procedure performed in reality. However, this test of the procedure does at least demonstrate that the procedure should work well in principle.

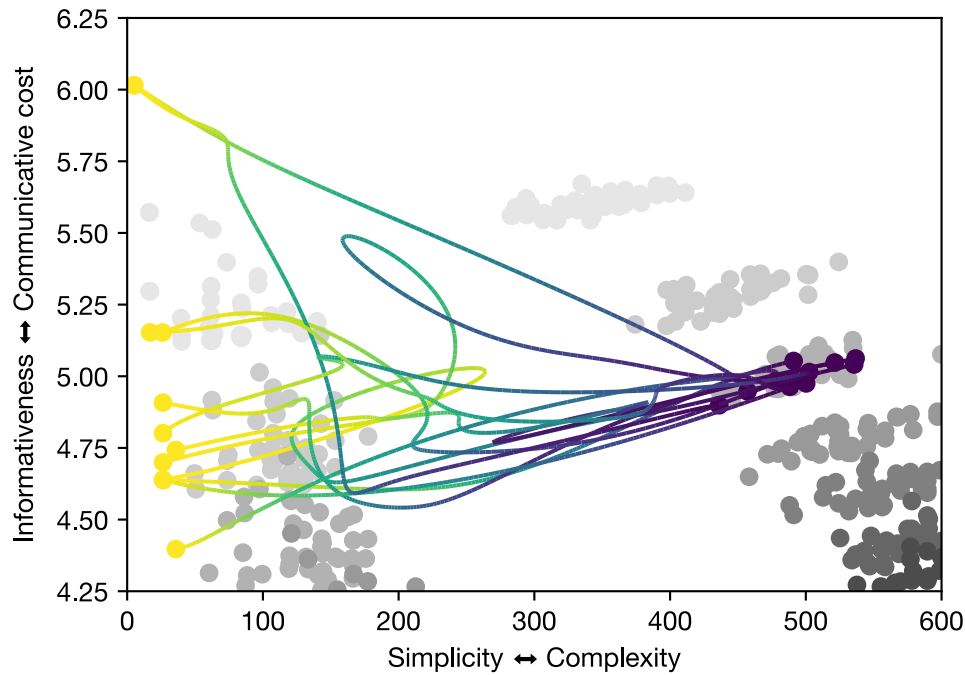
### 3.7 Conclusion to Chapter 3

This chapter provides computational and experimental evidence to show that learning is best understood through the lens of a simplicity principle and that, when the process of learning is iterated, category structure becomes increasingly simple, ultimately degenerating into its simplest possible form, the trivial partition (see Section 2.2.6). Iterated learning amounts to a westward expansion through simplicity–informativeness space as depicted in Fig. 3.10, which shows the evolutionary trajectories of the 12 iterated learning chains in Experiment 2 (cf. Fig. 2.14 on page 47). The languages become increasingly simple over time and only become informative to the extent that communicative cost can pick up on the emergent compactness.

The primary contribution of this project is to eliminate one of three possible explanations for convex/compact category structure. The three possibilities are:

1. Language users could discover through interaction that compact structures minimize error and therefore switch to using such systems, as suggested by Jäger and van Rooij (2007) and Gärdenfors (2014).
2. Language users could have a cognitive bias for informativeness, as argued by Fedzechkina et al. (2012) and Frank and Goodman (2014); assuming this bias is sufficiently sensitive to the compactness property, compact structure is then emergent from iterated learning, as suggested by Carstensen et al. (2015).
3. Language users could possess a simplicity bias, through a preference for either compressed representations (e.g. Gärdenfors, 2000; Sims, 2018) or compressible explanations (e.g. Culbertson & Kirby, 2016; Fass & Feldman, 2002), that is amplified by iterated learning, as shown in the paper.

Paper 2 rules out the second of these possibilities, but possibilities one and three remain; distinguishing between these two explanations could be the subject of future work, but



**Figure 3.10:** Evolutionary trajectories through simplicity–informativeness space for each of the 12 iterated learning chains. Each chain transitions from a purple dot (a random four-category language) to a yellow dot over the course of its evolutionary history, and the curves show smoothed trajectories through the space. The grey dots represent randomly generated systems that are convex (left) or random (right) with varying numbers of categories (darker grey = more categories). Together, the grey dots approximately delimit the space of possible languages. All chains become simpler by the final generation, pressing right up against the optimal frontier, and some chains become slightly more ‘informative’, although this is due the pressure for compactness from induction.

my own intuition is that the third option is the most promising.

This project has also made a number of methodological and practical contributions, including Python implementations of the Bayesian iterated learning model, the rectangle code, and the communicative cost framework, and JavaScript implementations of category and iterated learning experiments, all of which are available from <https://github.com/jwcarr/shepard>. These may prove useful to researchers working on related issues in the future.

Finally, despite our somewhat negative take on Regier and colleagues’ first foray into the language evolution literature (i.e. Carstensen et al., 2015), I hope that this project might still act as a bridge between these two bodies of work, which ultimately appear to be converging on very similar perspectives. In particular, I think it will be crucial to clearly define what is meant by the various jargon terms used by each.





## Chapter 4

# Informativeness from Interaction

Nun bewegt sich die Geschichte der Sprachen in der Diagonale zweier Kräfte: des Bequemlichkeitstriebes, der zur Abnutzung der Laute führt, und des Deutlichkeitstriebes, der jene Abnutzung nicht zur Zerstörung der Sprache ausarten lässt.<sup>16</sup>

— Georg von der Gabelentz (1891)

Von der Gabelentz was one of many early linguists to recognize that languages are shaped by competing forces (see also Martinet, 1952; Zipf, 1949), and the compromise he describes between *Bequemlichkeitsstreben* (striving for ease) and *Deutlichkeitsstreben* (striving for clarity) is akin to what I have called the simplicity–informativeness trade-off in this thesis (albeit in the domain of phonetics rather than semantics). The picture painted at the end of the last chapter was one of degeneration: In the limit, the iterated learning of category structure results in the trivial partition – a language that is so simple that it confers no useful benefit on its users. But as von der Gabelentz (1891, p. 251) notes, it is the desire for clarity – or informativeness – that prevents languages from degenerating entirely. Indeed, this is precisely what is argued by Kirby et al. (2015), who consider ‘degenerate’ languages as the product of learning, ‘holistic’ languages as the product of communication, and ‘compositional’ languages as the product of both pressures combined.

So, in this chapter, we turn to communicative interaction, which keeps the pressure

---

<sup>16</sup> Languages fluctuate in response to two opposing forces: The desire for ease, which leads to the erosion of sounds, and the desire for clarity, under which such erosion is not allowed to degenerate into the destruction of the language [my translation].

for simplicity from learning in check. In particular, this chapter demonstrates that a pressure for informativeness from communicative interaction is not merely a pressure for more categories (i.e. greater expressivity), but it also permits the emergence of a higher level form of structure, compositionality.

## 4.1 Preface to Paper 3

Paper 3 began life as my MSc project (Carr, 2013), supervised by Simon Kirby and Hannah Cornish. Following substantial additional work that I completed during my PhD, the paper was later published in *Cognitive Science* in May 2017 (Carr, Smith, Cornish, & Kirby, 2017). In total the paper represents about three months' work completed during my MSc (May to August 2013) and around 14 months' work completed during my PhD (primarily September 2014 to October 2015). The following parts of the paper were completed during my MSc:

1. The design and running of Experiments 1 and 2.
2. Parts of the Introduction and Methods sections, although most of these sections have been entirely revised.

The goal of my MSc project was to test Kirby et al.'s (2008) experimental paradigm under a meaning space in which there is no predefined categorical structure; instead, such categorical structure has to emerge alongside the evolution of compositional structure. We refer to this kind of space as *open-ended*, which is defined in the *Oxford Dictionary of English* as, 'having no predetermined limit or boundary; allowing the formulation of any answer, rather than a selection from a set of possible answers'. Experiments 1 and 2 map closely onto the two experiments in Kirby et al. (2008) – the second of which included an 'artificial expressivity pressure' to mimic the pressure from communication. In contrast to Kirby et al. (2008), we were not able to see the emergence of compositional structure under this artificial pressure.

Taking this work as a starting point, I developed Experiment 3 during my PhD which implemented a true pressure for informativeness – a communicative task. In addition, the methods used to analyse Experiment 1 and 2 were also entirely revised, which included running separate dissimilarity rating experiments. A particularly difficult aspect of the project was finding suitable methods for analysing the structure of

the emergent languages given that we do not actually know in advance what form that structure might take – that is for the participants to decide. In short, the following parts of the paper were completed during my PhD:

1. The design and running of Experiment 3.
2. The design and running of the online dissimilarity rating experiments.
3. Development of all analytical methods, which are almost entirely new over the approach taken in my MSc dissertation.
4. The writeup of the paper, including all figures and the supplementary material.
5. Revision of the paper following peer-review.
6. The additional material presented at the end of this chapter.

The paper is reproduced in full over the subsequent pages with the permission of the authors, who retain copyright. The footnotes may be found on page 136, and the citations may be looked up on pages 136–139 or in the references list at the end of this volume. The paper makes reference to three supplementary items, which may be found in the appendices at the end of this volume:

S1. *Experimental briefs*: Appendix D, page 189.

S2. *Geometric measure of triangle dissimilarity*: Appendix E, page 193.

S3. *MDS plots for all generations in all chains*: Appendix F, page 199.

All work reported in the paper is my own, including the technical development of the experiments and analytical methods. The contributions made by my coauthors were as follows:

**Kenny Smith** Advice on the design of Experiment 3 and the online dissimilarity rating tasks, advice on analytical methods, and general editing of the paper.

**Hannah Cornish** Advice on the design of Experiments 1 and 2.

**Simon Kirby** Advice on all three experiments and the online dissimilarity rating tasks, advice on analytical methods, and general editing of the paper.

# COGNITIVE SCIENCE

A Multidisciplinary Journal



Cognitive Science 41 (2017) 892–923

Copyright © 2016 The Authors. *Cognitive Science* published by Wiley Periodicals, Inc. on behalf of Cognitive Science Society.

All rights reserved.

ISSN: 0364-0213 print/1551-6709 online

DOI: 10.1111/cogs.12371

## The Cultural Evolution of Structured Languages in an Open-Ended, Continuous World

Jon W. Carr,<sup>a</sup> Kenny Smith,<sup>a</sup> Hannah Cornish,<sup>b</sup> Simon Kirby<sup>a</sup>

<sup>a</sup>*School of Philosophy, Psychology and Language Sciences, University of Edinburgh*

<sup>b</sup>*Psychology, School of Natural Sciences, University of Stirling*

Received 5 October 2015; received in revised form 11 January 2016; accepted 15 January 2016

### Abstract

Language maps signals onto meanings through the use of two distinct types of structure. First, the space of meanings is discretized into categories that are shared by all users of the language. Second, the signals employed by the language are compositional: The meaning of the whole is a function of its parts and the way in which those parts are combined. In three iterated learning experiments using a vast, continuous, open-ended meaning space, we explore the conditions under which both structured categories and structured signals emerge ex nihilo. While previous experiments have been limited to either categorical structure in meanings or compositional structure in signals, these experiments demonstrate that when the meaning space lacks clear preexisting boundaries, more subtle morphological structure that lacks straightforward compositionality—as found in natural languages—may evolve as a solution to joint pressures from learning and communication.

**Keywords:** Categorization; Communication; Compositionality; Cultural evolution; Iterated learning; Language evolution; Sound symbolism

### 1. Introduction

Language facilitates the division of the world into discrete, arbitrary categories (Lupyan, Rakison, & McClelland, 2007). For example, the words *bottle*, *cup*, *flask*, *glass*,

Correspondence should be sent to Jon Carr, School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Dugald Stewart Building, 3 Charles Street, Edinburgh, EH8 9AD, UK. E-mail: j.w.carr@ed.ac.uk

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

and *mug* separate the space of drinking vessels into discrete regions based on such features as shape, material, and function; however, languages differ in the way they discretize our continuous sensory perception of the observable world (Malt, Sloman, & Gennari, 2003). The presence of categorical structure in language reduces an intractable, theoretically infinite set of meanings to a tractable, finite set of words that have the flexibility to handle novel exemplars (Lakoff, 1987). By aligning on a particular system of categorical meaning distinctions, members of a linguistic population can rely on their shared understanding of the structure of the world to successfully communicate.

A second important property of language is its compositional structure: The meaning of a sentence—at multiple levels of analysis—is a function of the meanings of its parts and the way in which those parts are combined. For example, the meanings of *the water is in the cup* and *the cup is in the water* are predictable from the constituent parts (six monomorphemic words) and the word order. In language, compositional structure is a means for optimizing the trade-off between expressivity (the number of meanings that can be expressed) and compressibility (the degree to which the language can be reduced to atomic units and rules of recombination) (Kirby, Tamariz, Cornish, & Smith, 2015).

This paper focuses on how these two structural properties of language (categorical and compositional structure) can emerge simultaneously through the cultural evolutionary processes that are argued to hold at least some explanatory power in understanding where such structure comes from (e.g., Christiansen & Chater, 2008). Although the cultural evolution of categorical (e.g., Xu, Dowman, & Griffiths, 2013) and compositional (e.g., Kirby, Cornish, & Smith, 2008) structure has previously been demonstrated in isolation, we show here that structured languages can evolve where no categories have been provided by the experimenter a priori. We show this using an *open-ended* meaning space and the experimental paradigm of *iterated learning*.

### 1.1. Iterated learning

Iterated learning refers to “a process in which an individual acquires a behavior by observing a similar behavior in another individual who acquired it in the same way” (Kirby et al., 2008, p. 10681). For example, an individual learns a language from his or her parents, who themselves learned the language from their own parents. Taking inspiration from earlier computational (e.g., Hurford, 1989; Kirby, 2002; Smith, 2004) and experimental (e.g., Galantucci, 2005; Horner, Whiten, Flynn, & de Waal, 2006; Selten & Warglien, 2007) studies, Kirby et al. (2008) devised an experimental paradigm for studying iterated learning using adult human learners.

The basic design of an iterated learning experiment is as follows. An artificial language (i.e., a mapping between signals and meanings) is generated. In the case of Kirby et al. (2008), this language was a set of 27 randomly generated strings that were mapped onto a fixed set of 27 meanings (three shapes, in one of three colors, moving in one of three distinct patterns). Participants learn this language in a training phase and are then asked to reproduce the language by typing in the corresponding strings for a selection of meanings. The output from this test phase is then taught to a new participant, whose test out-

put is, in turn, taught to another new participant. These experiments typically show that, after several generations, the languages that initially started out as random evolve some form of structure.

The simplest kind of structure that can arise from these experiments is where participants collapse all meaning distinctions. This kind of language (referred to as “degenerate” by Kirby et al., 2015) is highly learnable because a single word can be applied to any meaning. Similarly, systems of structure can arise where the meaning space is collapsed into a small number of categories, each labeled by a distinct word. These kinds of structure represent one way in which languages might adapt to become easier to learn and therefore reliably transmitted. However, while these kinds of language are highly compressible, they are not expressive (see Kirby et al., 2015, for more discussion of this trade-off).

The second experiment reported by Kirby et al. (2008) implemented a “filtering” system that removed duplicate strings from the training material taught to the next participant in a chain, such that the training language always consisted of a set of unique signals. This modification was intended as an analog of the pressure for expressivity that exists in natural languages. In this experiment, small sets of meaningful, recombinable units emerged corresponding to the dimensions of the meaning space. For example, labels for all blue stimuli began with *l-* and labels for all stimuli moving in a spiral motion ended with *-pilu*. By learning a handful of linguistic units and the rules for combining them, participants were able to generate a unique label for any possible meaning combination, including meanings they had not been taught during training.

### *1.2. Continuous meaning spaces*

Iterated learning experiments have typically relied on meaning spaces that are discrete, finite, low dimensional, and structured by the experimenter. Kirby (2007) has described such meaning spaces as fixed and monolithic (p. 256). For example, the meaning space used in Kirby et al. (2008), described above, is three dimensional with each dimension (color, shape, and motion) varying over three discrete qualities. To take another example, the space in Smith and Wonnacott (2010) has two discrete dimensions (animal and plurality) for a total of eight meanings.

More recently, iterated learning experiments have been conducted using continuous meaning spaces (see also work with continuous signal spaces by e.g., Verhoef, 2012). Xu et al. (2013) conducted an experiment where participants had to label a continuous color space using between two and six color terms according to condition. The way in which a participant discretized the space was then taught to a new participant in a chain. After 13 generations of cultural transmission, the structure of the space came to resemble the way in which color space is typically structured by languages recorded in the World Color Survey (Kay, Berlin, Maffi, Merrifield, & Cook, 2009). For example, in the three-term condition, the emergent systems discretized the space into dark, light, and red categories.

Perfors and Navarro (2014) used a meaning space of squares that could vary continuously in terms of color (white to black) and size (small to large). In one condition, there

was an abrupt change in the color, such that the stimuli could be categorized into two broad categories (light-colored squares and dark-colored squares); in another condition, there was an abrupt change in the size of the squares. Labels for these stimuli were then passed along a transmission chain of learners. In both conditions, the authors found that the structure of the emergent languages came to mirror the structure of the meaning space, primarily making color or size distinctions according to condition.

Silvey, Kirby, and Smith (2013) produced a continuous meaning space by randomly generating four seed polygons and then gradually morphing the polygons into each other, creating a space of 25 stimuli. The space had no obvious internal boundaries; as such, participants showed variation in how they discretized it. The authors also conducted an iterated learning experiment using the same meaning space (Silvey, 2014, Chapter 5). In this experiment, each generation consisted of a pair of participants who communicated about the stimuli using a fixed set of up to 30 words. Over five generations, the category systems that emerged tended to make fewer distinctions and became easier to learn. Furthermore, the category structures became increasingly convex, providing experimental evidence for predictions made by Gärdenfors (2000) about semantic convexity.<sup>1</sup>

### *1.3. Research questions*

Two important and related questions arise from prior research into iterated learning. First, to what extent are the general findings supported under more realistic assumptions about meaning? For example, do the results still hold when the meaning space possesses properties that more closely reflect the natural world (e.g., high-dimensionality, open-endedness, continuousness)? This question has been partially addressed by the work with continuous meaning spaces described above (see also simulation work by e.g., Laskowski, 2008). The second question that arises is whether iterated learning simply returns the structure prescribed by the experimenter, transferring it from one domain (e.g., predefined categories in the meaning space) to another domain (e.g., the emergent structure in the signals). Xu et al. (2013) address this issue to a certain extent; however, the participants in their experiment are explicitly told how many categories to create—the number of categories does not arise naturally—and the participants are also likely to have strong preconceptions of how to discretize color space based on the color system of their native language (although the authors do address this); furthermore, Xu et al. (2013) do not test for emergent signal structure, since a fixed set of labels is provided. If it is indeed the case that iterated learning experiments simply return structure provided by the experimenter, is it realistic to assume that structured languages can evolve in a context where individuals are not provided with shared categorizations of the observable world?

In this paper, we address these concerns by introducing a novel meaning space of randomly generated triangle stimuli. Like previous work, our meaning space is continuous, but crucially it is also open-ended: The structure of the space is neither provided by the experimenter nor naturally categorizable; instead it is up to the participants to arbitrarily decide how to categorize the space. In addition, the experiment is set up in such a way that no two generations are tested on or trained on precisely the same stimuli, forcing

participants to generalize from the training stimuli to the test stimuli in all cases. Finally, the space of possible stimuli that participants can encounter is vast, forcing participants to adopt a system of categorization. Together, these properties of our meaning space represent more realistic assumptions about the natural world, and by not defining what the meaning dimensions are, we can test whether structure can arise in the signals and in the meaning space simultaneously.

#### *1.4. Outline of this paper*

This paper reports three artificial language learning experiments that use the paradigm of experimental iterated learning described above. Experiment 1 (basic transmission) looks at what happens when there is no pressure for expressivity. It therefore provides a baseline for how participants respond to the open-ended meaning space. The results demonstrate that categories emerge over generational time to discretize the space of possible triangles. Experiment 2 (transmission with an artificial expressivity pressure) explores whether compositional structure can emerge alongside the categorization of the meaning space by implementing an artificial pressure for expressivity. The results of this experiment were negative, suggesting that the second experiment reported by Kirby et al. (2008) may be a special case relating to the discrete meaning space adopted therein. Experiment 3 (transmission with communication) implements a natural expressivity pressure—communication—and shows that sublexical structure can emerge when languages are both learned and used to communicate.

## **2. Experiment 1: Basic transmission**

Our first experiment is equivalent to the first experiment reported by Kirby et al. (2008) and looks at what happens when languages are passed along a simple transmission chain with no pressure for expressivity. We had two hypotheses about what would happen over generational time:

1. We expect that the languages will become increasingly easy to learn.
2. We expect to find emergent categories in the meaning space.

These outcomes were expected because the languages should adapt to the cognitive biases of the language users, gradually becoming more learnable. Categories are a way to increase learnability because they constitute a more compressed representation of the meaning space.

### *2.1. Method*

The experiment adopted the standard iterated learning paradigm described previously: Participants were arranged into transmission chains in which the output from generation  $i$  became the input to generation  $i + 1$  for a given chain.



### 2.1.1. Participants

Forty participants (20 female) were recruited at the University of Edinburgh. The median age was 22 years (range: 19–34). Participants were paid £5.50 for participation, and a £20 Amazon voucher was offered as a prize for the best learner. Ethical approval was granted for all experiments reported in this paper according to the procedures of the School of Philosophy, Psychology, and Language Sciences at the University of Edinburgh. All participants provided informed consent and were offered debrief information.

### 2.1.2. Stimuli

Participants learned and produced artificial languages that consisted of labels paired with triangles. To generate a triangle stimulus, three points were chosen at random in a 480×480-pixel space and joined together with black lines (2 pixels wide). The space was enclosed in a 500×500-pixel dashed, gray bounding box. One vertex (determined randomly) was marked with a black circle with a radius of eight pixels (referred to as the *orienting spot*). Its function is to give the participant some context about which way the triangle is oriented, although this was not explicitly explained to participants. The number of stimuli<sup>2</sup> that can be generated in this way is  $3 \binom{480^2}{3} \approx 6 \times 10^{15}$ . See Fig. 1 for some examples of the triangle stimuli. In this paper, we use the terms *dynamic set* and *static set* to refer to subsets from the set of possible triangles that participants may be exposed to. These terms are explained in greater detail below; for now it suffices to say that a unique dynamic set is generated at every generation (i.e., it changes across participants and generations), while the static set is identical for all participants across all experiments, allowing us to take measurements on a consistent set of stimuli.

The labels used as input to the first generation in a chain were generated by concatenating 2–4 syllables at random. A syllable consisted of a consonant from the set  $\{d, f, k, m, p, z\}$  and a vowel from the set  $\{a, i, o, u\}$  (pronounced /a i ou u/), yielding 24 possible syllables. The labels used as input to subsequent generations were derived from the output of the previous generation in the chain. We used the MacinTalk speech synthesizer (Alex voice) to produce a synthesized spoken version of each label with primary stress

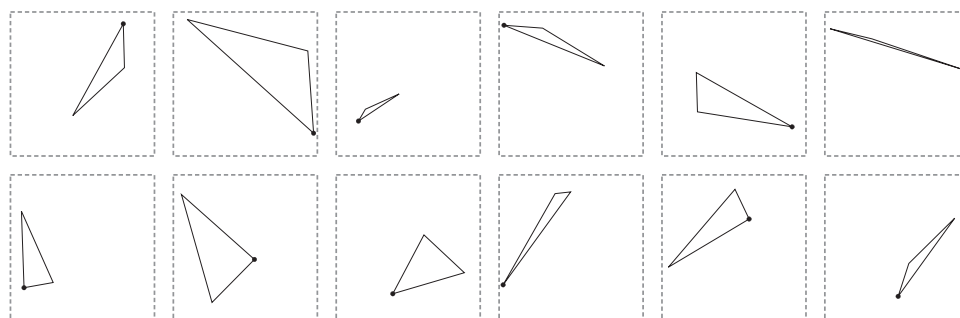


Fig. 1. Examples of the triangle stimuli. The stimuli are generated by randomly selecting three points inside a dashed, gray bounding box. One vertex is marked with a black circle.

on the penultimate syllable. The use of spoken stimuli, alongside the written stimuli, offers a number of benefits: (a) it makes the task more engaging, (b) it frees participants from having to consider how to pronounce or subvocalize the words, (c) it ensures that all participants hear the words pronounced in the same way, and (d) it ensures that participants still hear the word even if they only pay attention to the triangle stimulus and ignore the written label. When participants introduced new characters, those characters were assigned phonological values consistent with English orthography.

### 2.1.3. Procedure

Participants were assigned to one of four chains at random until the chain reached 10 generations. Participants were told that they would be learning the language of the *Flatlanders* (after Abbott, 1884), a fictional life-form that has many words for triangles. The task was explained to participants in a written brief (see Appendix S1 in the supplementary material), the contents of which were reiterated verbally. The experiment was divided into a training phase followed by a test phase. The training phase involved learning the labels used by the previous participant. The test phase involved providing labels for novel triangles. The experimental procedure is illustrated in Fig. 2, and each phase is explained in the following paragraphs.

During training, participants learned the labels that the previous participant had applied to the 48 triangles in his or her dynamic set (i.e., the unique set of stimuli generated for the previous participant's test phase). Each training trial lasted 5 s. On each trial, the triangle was presented first, and its associated label appeared below it after a 1 s delay to ensure that both stimuli were attended to. The synthesized form of the label was played through headphones at the same time as the presentation of the written form. Training was done in three blocks. In each block, the participant was exposed to the 48 items in a

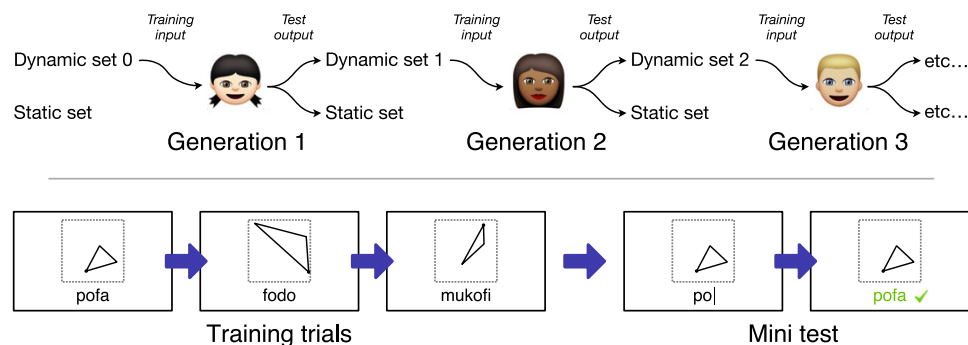


Fig. 2. (Top) The participant at generation  $i$  is trained on a set of triangle stimuli paired with labels (dynamic set  $i - 1$ ). He or she is then tested on two novel sets of triangles: a randomly generated set (dynamic set  $i$ ) and a set that remains constant for all participants (the static set). The labels applied to the dynamic set become the training input to generation  $i + 1$ . (Bottom) During training, the participant sees a series of three triangles along with their associated labels. One of the three triangles is then presented again, and the participant is prompted to type its associated label. Feedback is then given on whether the answer was correct.

randomized order for a total of 144 trials. After every third trial (i.e., 16 times per block, 48 times overall), the participant was shown one of the previous three triangle stimuli again and prompted to type its label. We refer to this as a *mini test*. Over the course of training, each of the 48 items was mini-tested once. Feedback on each mini test was given in the form of a green checkmark or a red cross according to whether the participant answered correctly. If the answer was incorrect, the correct answer was shown. The mini tests were intended as a means for holding the participant's attention during the training phase.

In the test phase, participants were exposed to 96 triangle stimuli, none of which they had seen during training, and were prompted to type the associated label for each one. The 96 stimuli consisted of the 48 stimuli in a newly generated dynamic set (which would go on to become the training material for the subsequent participant in the chain) and the 48 stimuli in the static set (in randomized order). The presentation of these two sets was interleaved. The static set comprised the same set of triangles across all participants in all experiments, allowing us to take measurements on a consistent set of stimuli. No feedback was provided during the test phase, since there is no right or wrong answer.

## 2.2. Results

The results for Experiment 1 are shown in Fig. 3 and are discussed in the following sections. The raw data and analysis are available from <https://github.com/jwcarr/flatlanders>.

### 2.2.1. Loss of expressivity

We can estimate how expressive a language is by looking at the number of words it contains. A language with more words is potentially capable of making more meaning distinctions. In the initial Generation-0 input, 48 unique strings were used to label the static set, but by Generation 10, this number decreased to 6 or 7, and in Chain D, a single word, *mika*, was used to describe all triangles. These results are shown in Fig. 3A. Page's test (Page, 1963) revealed that this decrease in the number of unique labels was significant ( $L = 1,993$ ,  $m = 4$ ,  $n = 11$ ,  $p < .001$ ). These results show that the languages are becoming less expressive over time.

### 2.2.2. Increase in learnability

We expected to find that the languages would become increasingly learnable over time. If a language is easy to learn, a participant's output language should more faithfully reproduce the rules of the input language. In other words, we would expect to find a decrease in intergenerational transmission error over time. Intergenerational transmission error was measured by taking the mean normalized Levenshtein edit-distance<sup>3</sup> (Levenshtein, 1966) between the strings used to describe items in the static set at generation  $i$  and the corresponding strings at generation  $i - 1$ :

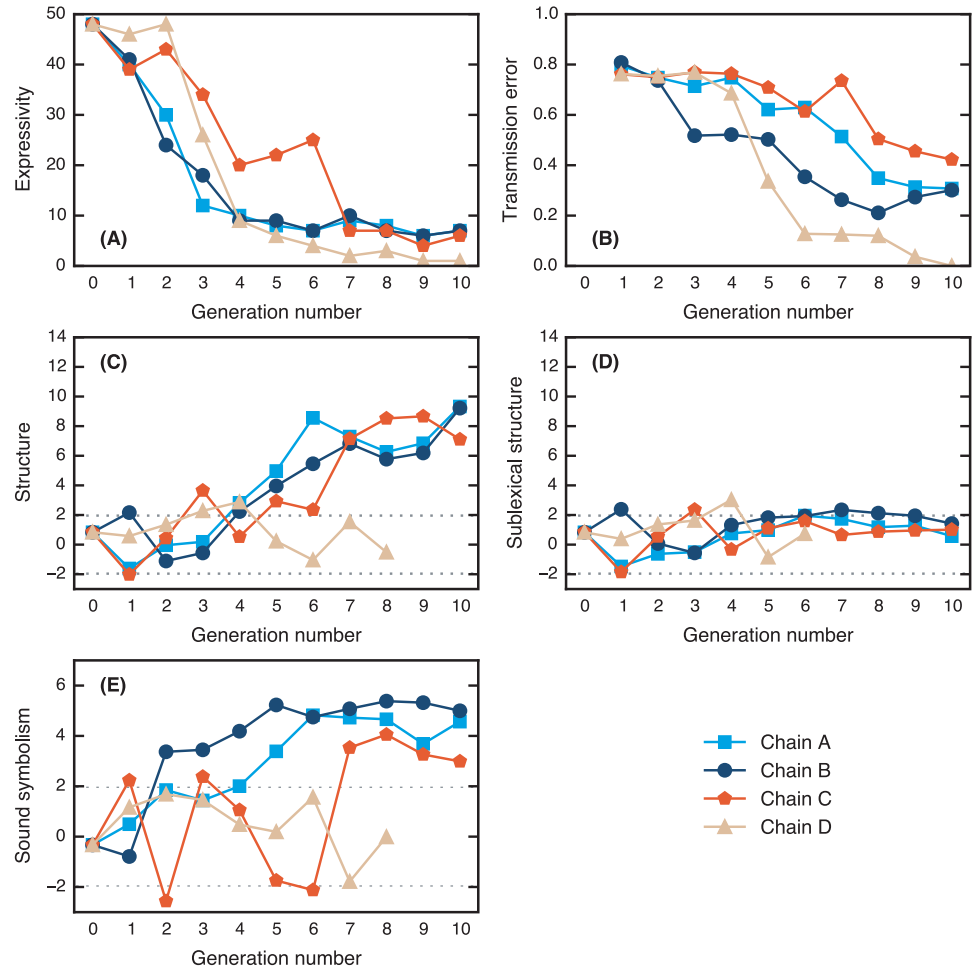


Fig. 3. Results of Experiment 1. (A) Expressivity: number of unique strings in the static set. (B) Levels of transmission error. (C) Levels of general structure. (D) Levels of sublexical structure. (E) Levels of shape-based sound symbolism. The dotted lines in (C), (D), and (E) give the upper and lower 95% significance levels; points lying outside of this interval are unlikely to be explained by chance. Some data points at the end of Chain D are undefined due to the small number of unique strings.

$$\frac{1}{48} \sum_{m=1}^{48} \frac{\text{LD}(s_i^m, s_{i-1}^m)}{\max[\text{len}(s_i^m), \text{len}(s_{i-1}^m)]}, \quad (1)$$

where LD gives the Levenshtein edit-distance,  $s$  is a string, and  $m$  is a meaning from the static set of 48 items. This measure of error is expressed in  $[0, 1]$ , where 0 is perfect alignment between consecutive generations. The results for transmission error are shown

in Fig. 3B. Page's test revealed that the decrease in transmission error was significant ( $L = 1,514$ ,  $m = 4$ ,  $n = 10$ ,  $p < .001$ ), suggesting that the languages are becoming easier to learn over time. Although transmission error may appear quite high by the final generation, this should not be surprising, since a score of 0 requires not only that consecutive participants label the categories in the same way, but also that they infer the same category boundaries; in natural languages, however, the boundaries between categories are known to be fuzzy (Rosch, 1973).

### 2.2.3. Emergence of structure

Although the languages became less expressive, we expected to find that the words would increasingly be used to categorize the space systematically. In a systematic language, we would expect to find that similar labels refer to similar meanings, while dissimilar labels refer to dissimilar meanings. Thus, to measure how structured the system is, we correlate the dissimilarity between pairs of strings with the dissimilarity between pairs of triangles for all  $n(n - 1)/2$  pairs. The normalized Levenshtein edit-distance was used as a measure of dissimilarity between strings. To measure the dissimilarity between triangles, we conducted a separate experiment in which naïve participants were asked to rate the dissimilarity between pairs of triangles (see Appendix A for full details of this experiment and Appendix S2 in the supplementary material for an alternative geometric approach). Following previous studies (e.g., Kirby et al., 2008, 2015), the distance matrices for string dissimilarity and triangle dissimilarity were correlated using the Mantel test (Mantel, 1967), since the distances are not independent of each other making standard parametric statistics unsuitable. The test compares the Pearson correlation for the veridical signal–meaning mapping against a distribution of Pearson correlations for permutations of the mapping, yielding a standard score ( $z$ -score). The results of this analysis are presented in Fig. 3C. The last two generations of Chain D are undefined under this measure because there is only one word in the language. The plot shows that structure is emerging in all chains with the exception of Chain D. Page's test revealed a significant increase in structure ( $L = 1472$ ,  $m = 3$ ,  $n = 11$ ,  $p < .001$ ; excluding Chain D due to missing data points).

However, this measure of structure cannot discriminate between category structure and string-internal structure (e.g., compositionality). To test if structure was present inside the signals, a modification was made to the measure: Rather than randomize the mapping between signals and meanings, we randomize the mapping between the category labels (i.e., the unique set of words in the language) and the sets of triangles they map onto, such that the set of triangles labeled by a given word remains intact but the labels for each category are randomly shuffled. Under this randomization method, any categorical structure in the language remains present in the permuted mappings, so a high  $z$ -score indicates that there must be additional structure present inside the strings themselves. The results from this alternative approach are shown in Fig. 3D, where the majority of data points are below the upper 95% significance level, suggesting that there is no string-internal structure in the languages of this experiment.

To visualize the categories, the pairwise dissimilarity ratings (obtained from the naïve raters; Appendix A) were passed through a multidimensional scaling (MDS) algorithm, producing a two-dimensional representation of the meaning space.<sup>4</sup> MDS finds an arrangement of items in a metric space that best preserves the distances known to exist between those items (see e.g., Borg & Groenen, 2005). The MDS solution is shown in the plot in Fig. 4. Each dark dot represents one of the triangles in the static set; triangles that are close together in this space were rated to be similar, and triangles that are far apart were rated to be dissimilar. Although the dimensions of the space are somewhat abstract, the *x*-axis appears to correspond to shape, while the *y*-axis appears to have some correspondence with size. The space is partitioned into 48 Voronoi cells—one cell for each triangle in the static set. Each cell encompasses all points in the space that lie closer to the associated triangle than to any other triangle from the static set. In other words, each Voronoi cell delimits the space of triangles that would have been labeled with the associated string under the assumption that each item is a prototypical member of a convex category (Gärdenfors, 2000).

Color is used in Fig. 4 to show information about the state of the language at Generation 10 in Chain A; similarity in color indicates similarity in word form. To determine a

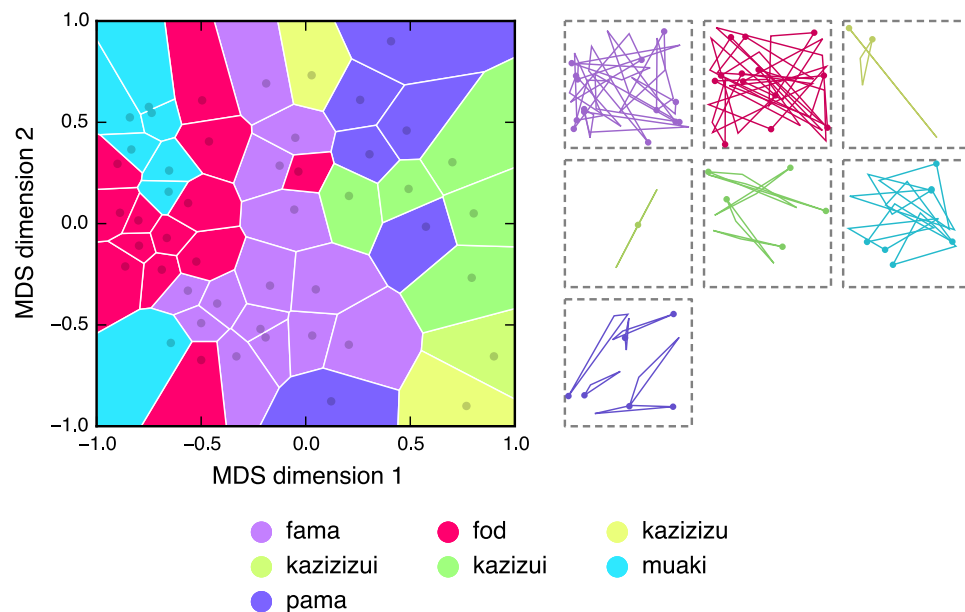


Fig. 4. Categorical structure of the meaning space at Generation 10 in Chain A. The plot on the left shows how the meaning space is discretized by the words in the language: Similarity in position represents similarity in meaning; similarity in color represents similarity in word form. On the right, all triangles in the static set are grouped by the word used to describe them (presented in the same order as the legend). Refer to the main text for a full description and interpretation of this figure.

color for each word, we computed the pairwise Levenshtein edit-distances between the seven words in this particular language and derived a two-dimensional MDS solution centered on the origin. The Cartesian coordinates in this MDS space were converted to polar coordinates and then mapped into HSV (hue, saturation, intensity value) colorimetric space: The angular coordinate was mapped to hue and the radial coordinate (scaled in  $[.5, 1]$  to avoid overly dark colors) was mapped to saturation; the intensity value was held constant at 1 (see Lespinats & Fertil, 2011, for a full description of this method). The seven words are given in the legend alongside their assigned colors. Each Voronoi cell is colored according to the word that was used to describe its associated triangle, making it possible to see how the space is discretized by the words. The plot is a visual approximation of the measure of structure described above: In a structured language, similar colors will cluster into similar regions, while in an unstructured language, colors will be randomly distributed across the space. The images to the right of the plot show all triangle stimuli in the static set grouped and colored according to the word that was used to describe them. Note that Fig. 4 combines two data sets: The structure of the meaning space is determined by the naïve raters, while the color coding is determined by how the participant at Generation 10 in Chain A labeled the triangles. Figures for all generations in all chains can be found in Appendix S3 in the supplementary material.

Fig. 4 clearly shows that the language divides the meaning space into around five categorical regions. The center of the space (medium thin triangles) is occupied by the word *fama* (light purple), with the similar word *pama* (dark purple) branching off into the top-right corner (smaller thin triangles). The *kazi-* forms (*kazizizu*, *kazizizui*, and *kazizui*; yellow-green) occupy the right-hand side of the plot and represent the extremely thin triangles. *Muaki* (blue) mostly occupies the top left (smaller open triangles), and *fod* (pink) occupies the center left (larger open triangles). With some exceptions, the five main categories tend to form single, contiguous regions (e.g., it is possible to travel between any two examples of a *fama* without leaving the *fama* region), although the regions do not appear to be convex (it is not always possible to travel in a straight line without passing through another category). It is important to note, however, that the Voronoi tessellation of MDS space only offers a two-dimensional model of participants' underlying conceptual representations of the triangles and linguistic categories; the plots should therefore not be taken as a reliable source of information about the precise structuring of the meaning space.

#### 2.2.4. The rise of sound-symbolic languages

Sound symbolism describes the phenomenon where a unit of sound goes “beyond its linguistic function as a contrastive, non-meaning-bearing unit, to directly express some kind of meaning” (Nuckolls, 1999, p. 228). Although we did not initially set out to test for the emergence of sound-symbolic languages, it appeared that such patterning might be present. For example, the word *kiki* (the same word used in the classic bouba/kiki experiments; Köhler, 1929) arose independently in several chains (Chains C and D in this experiment and Chains E, G, and H in Experiment 2) to describe very thin or small triangles. To explore the emergence of shape-based sound symbolism, we hypothesized that



the extent to which each triangle was thin vs. equilateral would be correlated with the presence of phonemes associated with pointy vs. round stimuli (following e.g., Köhler, 1929; Kovic, Plunkett, & Westermann, 2010; Maurer, Pathman, & Mondloch, 2006). The “equilateralness” of a triangle (a proxy for shape) was calculated as

$$\frac{a}{p^2/(12\sqrt{3})}, \quad (2)$$

where  $a$  is the triangle’s area and  $p$  is its perimeter.<sup>5</sup> To measure the “roundedness” of a string, we used the sound-symbolic correspondences described by Ahlner and Zlatev (2010, p. 310) to divide all phonemes that occurred into three categories: “round” phonemes /b d g l m n oʊ ɔ u/, which received a score of +1, “pointy” phonemes /k p t ei i/, which received a score of −1, and all other phonemes, which received a score of 0. We then correlated the total roundedness of the strings with the equilateralness of the corresponding triangles and compared this correlation to a distribution of correlations for permutations of the mapping between signal and meaning to arrive at a standardized measure of shape-based sound symbolism. The results are shown in Fig. 3E; by the final generations, there are significant levels of shape-based sound symbolism in chains A, B, and C.

The same analysis was conducted for size-based sound symbolism using the centroid size<sup>6</sup> as a measure of a triangle’s size. This measure is uncorrelated with the triangle’s shape (Bookstein, 1991, p. 97), which is particularly important given the great amount of overlap in phonemes associated with both shape and size. Specifically, the “bigness” of a string was measured based on the phonemes listed in Thompson and Estes (2011, p. 2396): The “big” phonemes /b d g l m w a oʊ ɔ u/ received a score of +1 and the “small” phonemes /k p t ei i/ received a score of −1. While there was an effect in some later generations, the results were quite weak. Given the lack of a strong effect for size, only the shape-based sound symbolism results are reported in this paper.

#### 2.2.5. Summary of Experiment 1

The results for Experiment 1 suggest that categorical structure emerges in the languages. In Chains A, B, and C, the space of possible triangles was gradually divided into a small number of arbitrary categories that varied across chains. In Chain D, a single word came to stand for all triangles, which is itself a form of categorical structure—in everyday English, for example, all three-sided, two-dimensional figures can be categorized under the single word *triangle*. The small number of words that emerged in the languages by the final generations mirrors the underspecification found in the first experiment of Kirby et al. (2008). Categories allow for languages that are more compressed and, as such, more learnable. For example, the language depicted in Fig. 4 can be minimally represented by seven words, but it is presumably capable of describing any of the  $6 \times 10^{15}$  triangles that could have been generated. However, highly compressed languages are not necessarily useful in the context of language use, where it is important to be able to disambiguate one referent from a set of referents (see Kemp & Regier, 2012, for an example of this trade-off in the



context of kinship categories). To test whether more expressive languages could evolve under this unstructured, open-ended meaning space, we conducted two additional experiments that include expressivity pressures.

### 3. Experiment 2: Transmission with an artificial expressivity pressure

Our second experiment tests whether artificially forcing participants to use expressive languages results in compositional structure as a solution to maintaining both diversity of forms and compressible (and therefore learnable) languages. We had three hypotheses:

1. We expect that the languages will become increasingly easy to learn.
2. We expect to find emergent categories in the meaning space.
3. We expect to find emergent structure in the signals (e.g., compositionality).

The addition of Hypothesis 3 to the two hypotheses of Experiment 1 was motivated by Kirby et al. (2008), whose second experiment showed that forcing languages to remain expressive results in emergent compositional structure. In our experiment, participants could, for example, use a system where the first syllable (*a*, *b*, or *c*) denotes three sizes, the second syllable (*d* or *e*) denotes broad or thin, and the third syllable (*f*, *g*, *h*, or *i*) denotes the quadrant that the triangle is primarily located in. In this example, participants would only need to learn nine linguistic units (syllables *a–i*) and the rules for combining them but would be able to generate  $3 \times 2 \times 4 = 24$  distinct words, providing referential precision at minimal cost in terms of the number of label components to be learned.

#### 3.1. Method

##### 3.1.1. Participants

Forty participants (25 female), none of whom took part in Experiment 1, were recruited at the University of Edinburgh. The median age was 22 years (range: 18–50). Participants were paid £5.50 for participation, and a £20 Amazon voucher was awarded to the best learner.

##### 3.1.2. Procedure

The procedure was identical to Experiment 1, except that participants could not use the same string more than three times to label test items from the dynamic set (i.e., every other test trial). We did not impose this limitation on the static set because only the dynamic set can lead to a runaway loss of expressivity, since the way in which this set was labeled would be passed to the next generation. The advantage of this approach is that participants will only encounter the expressivity pressure in half of trials. The disadvantage is that the static set may not be entirely representative of how the participant responded in the dynamic set. In dynamic set trials, upon attempting to enter a word that had previously been used three times, the participant was presented with the message “You’ve used this word too often. Please use another word.” An additional sentence was

added to the brief to explain that this could happen (see Appendix S1 in the supplementary material). This modification to the test procedure forces the languages to remain expressive, since the output languages passed to the next generation must contain a minimum of  $48 / 3 = 16$  unique strings.

### 3.2. Results

The results of Experiment 2 are shown in Fig. 5 and are discussed in the following sections.

#### 3.2.1. Expressivity

The number of unique strings used to label items in the dynamic set was not able to collapse so dramatically. Although the pressure was only applied to the dynamic set, the number of unique strings in the static set also remained high (as shown in Fig. 5A). The languages thus remain more expressive than Experiment 1.

#### 3.2.2. Learnability

Fig. 5B shows that intergenerational transmission error in Experiment 2 remained relatively static. Nevertheless, the results do show a significant decrease ( $L = 1,415$ ,  $m = 4$ ,  $n = 10$ ,  $p < .001$ ) from an average of 80% error at Generation 1 down to an average of 66% error at Generation 10.

#### 3.2.3. Structure

Although the languages in Experiment 2 are more expressive, this did not translate into increased levels of structure. Like Experiment 1, there is no evidence for sublexical structure (Fig. 5D); however, levels of general structure are also low (Fig. 5C), with only Chains G and H showing marginal, albeit fragile, levels of structure. Fig. 6 shows the state of the language at Generation 8 in Chain G. In this example, which was the most structured language to emerge, there is a clear tendency for similar labels to cluster together. For example, labels colored green cluster down the right-hand side, dark blues in the top left, orange–yellows on the left-hand side, and so forth. However, the structure of the space is not as clear cut as in the case of Experiment 1, partly due to the increased number of words. In general, however, strong levels of categorical structure did not develop in this experiment (as indicated by Fig. 5C), and it seems that the participants continue to make a small number of categorical distinctions by using similar (but not necessarily identical) strings to label each category. For example, although the language shown in Fig. 6 uses 14 labels, there appear to be five broad categories (colored blue/cyan, green, magenta, orange/yellow, red/salmon; this is not simply an artifact of color perception as these five broad categories are also clear from the strings themselves).

#### 3.2.4. Sound symbolism

Like Experiment 1, there are significant levels of shape-based sound symbolism emerging in some of the later generations (Fig. 5E), although the effect tends to be weaker.

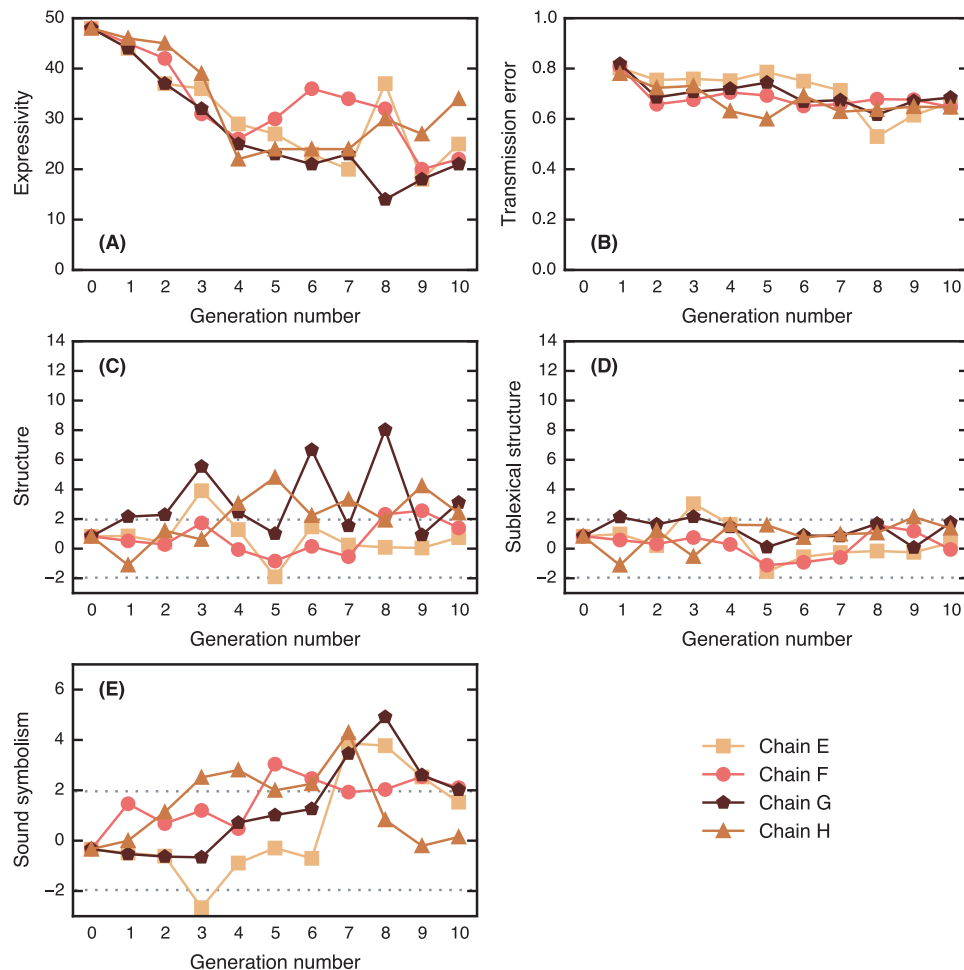


Fig. 5. Results of Experiment 2. (A) Expressivity: number of unique strings in the static set. (B) Levels of transmission error. (C) Levels of general structure. (D) Levels of sublexical structure. (E) Levels of shape-based sound symbolism. The dotted lines in (C), (D), and (E) give the upper and lower 95% significance levels; points lying outside of this interval are unlikely to be explained by chance.

### 3.2.5. Summary of Experiment 2

Placing a limit on the number of times a particular word could be reused allowed the languages to remain expressive. However, this did not translate into compositional structure as hypothesized. In fact, the substantial variation in the languages prevented many of the participants from stabilizing on a set of reliable categories. This result is at odds with the second experiment reported by Kirby et al. (2008), where an artificial pressure was

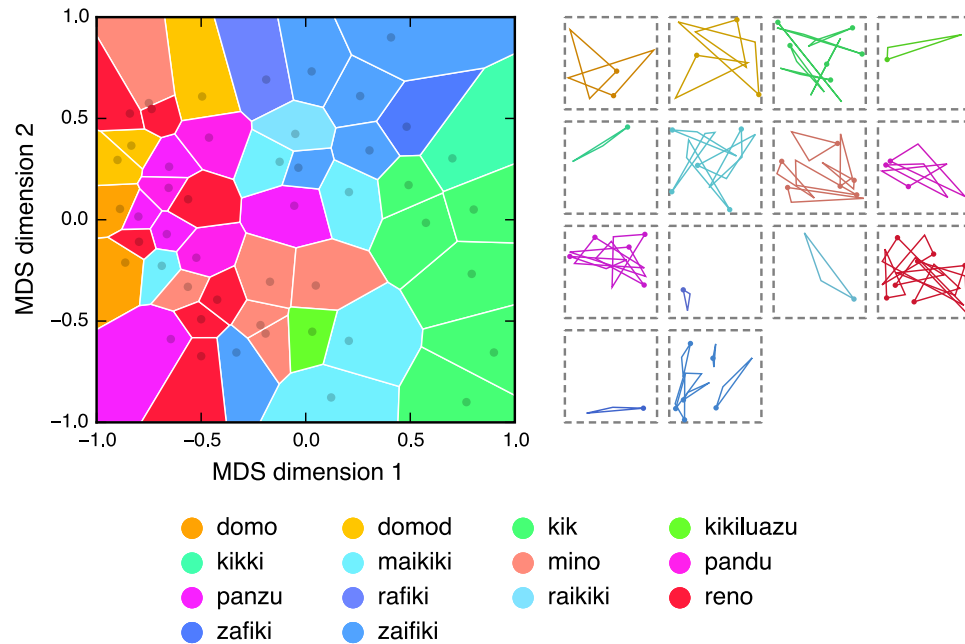


Fig. 6. Categorical structure of the meaning space at Generation 8 in Chain G. The plot on the left shows how the meaning space is discretized by the words in the language: Similarity in position represents similarity in meaning; similarity in color represents similarity in word form. On the right, all triangles in the static set are grouped by the word used to describe them.

sufficient to give rise to compositional languages. While there are many possible explanations for this, one possibility is that an artificial pressure for expressivity is only sufficient in the artificial case of a small, discrete, structured meaning space.

#### 4. Experiment 3: Transmission with communication

The restriction imposed on Experiment 2 was artificial; although participants had to remain expressive, there was no natural reason to use a large number of distinct strings. In our final experiment, we replaced the artificial expressivity pressure with a more ecologically valid pressure: At each generation, two participants must use the language to communicate with each other. Communication introduces a natural pressure for expressivity because, in order to maximize their communicative success, a pair of participants will need a language that is well-adapted to the discrimination of referents in a world of triangles. Our hypotheses were identical to those of Experiment 2.

#### 4.1. Method

##### 4.1.1. Participants

Eighty participants (63 female) were recruited at the University of Edinburgh, none of whom took part in Experiments 1 or 2. The median age was 21 years (range: 18–37). Participants were paid £8.50 for participation. The pair of participants who were most successful at communicating were both awarded a £20 Amazon voucher to encourage participants to be as communicative as possible with their partners.

##### 4.1.2. Procedure

The task was explained to participants in a written brief (see Appendix S1 in the supplementary material), the contents of which were reiterated verbally. The procedure followed the same communication game paradigm introduced in other iterated learning experiments (e.g., Kirby et al., 2015; Winters, Kirby, & Smith, 2015); this is illustrated in Fig. 7. Sitting in separate booths, a pair of participants completed the same training regimen used in Experiments 1 and 2. The training material presented to the two participants was identical and was derived from the dynamic set of the previous generation. Once both participants had completed training, they entered a communication game in which they took turns to play the role of director and matcher. The director was shown a triangle stimulus on his or her screen and was asked to describe that triangle to his or her partner. This label was then displayed on the matcher's screen along with six triangles to

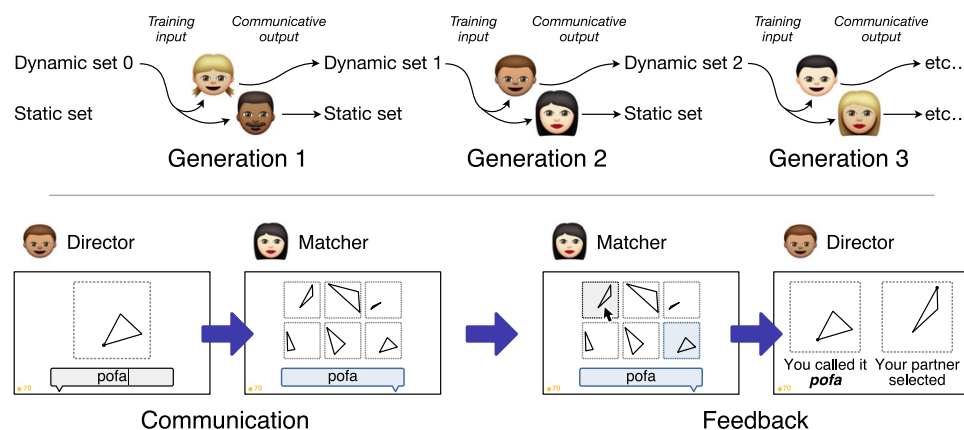


Fig. 7. (Top) The participants at generation  $i$  are individually trained on dynamic set  $i - 1$ . They then communicate about two novel sets of triangles: a randomly generated set (dynamic set  $i$ ) and a set that remains constant for all participants (the static set). The labels applied to the dynamic set become the training input to generation  $i + 1$ . (Bottom) During communication, the director is shown a triangle and is prompted to type a label to describe it. The label is then displayed on the matcher's screen along with an array of six triangles to choose from. The matcher's task is to click on the triangle that his or her partner is trying to communicate. As feedback, both participants see the target triangle and the selected triangle.

choose from (the context array). The context array contained the target triangle (in randomized position) and five randomly generated distractors. The matcher's task was to click on the triangle that his or her partner was trying to communicate. The director and matcher were provided with full feedback: After making a selection, the correct target in the context array was highlighted in blue, and the director was shown the triangle that the matcher had selected alongside the correct target. The participants were jointly awarded 10 points for each correct match; the number of points accumulated was shown in the bottom left corner of both screens throughout the communication game.

One of the participants (determined randomly) labeled the dynamic set and the other labeled the static set for a total of 96 communication trials. Like the previous experiments, the dynamic and static sets were labeled in alternation as the pair of participants swapped roles. This approach means that the subsequent generation was exposed to input from one cultural parent (the participant who labeled the dynamic set); the disadvantage is that the static set is only representative of the participant who labeled that set.

## 4.2. Results

The results of Experiment 3 are shown in Fig. 8 and are discussed in the following sections.

### 4.2.1. Expressivity

The expressivity results are shown in Fig. 8A. The number of unique strings is generally greater than that observed in Experiment 1, and the number of unique strings in Chain J and the first half of Chain L is comparable to Experiment 2.

### 4.2.2. Learnability

The results for transmission error are shown in Fig. 8B. There is a significant decrease ( $L = 1,503$ ,  $m = 4$ ,  $n = 10$ ,  $p < .001$ ) from an average of 80% error at Generation 1 down to an average of 50% error at Generation 10.

### 4.2.3. Communicative accuracy

Fig. 8C shows the number of times the communicating pair correctly identified the target triangle out of 96 trials. The chance level of accuracy under this measure is  $96 / 6 = 16$  (indicated by the dotted line). All but one of the pairs scored above chance. There was a significant increase ( $L = 1,321.5$ ,  $m = 4$ ,  $n = 10$ ,  $p = .021$ ), with later generations tending to make more correct matches. Fig. 8D shows a more fine-grained measure of communicative accuracy: the total dissimilarity between the selected triangle and the target triangle for all incorrect responses (dissimilarity scores were collected in a separate experiment; see Appendix B). This gives a measure of the amount of communicative error at each generation. There was a significant decrease ( $L = 1,356$ ,  $m = 4$ ,  $n = 10$ ,  $p = .004$ ), which again indicates that later generations communicate more accurately. Nevertheless, levels of communicative accuracy were quite low. The pair of participants with the highest score was Generation 8 in Chain J (46 correct trials). That all partici-

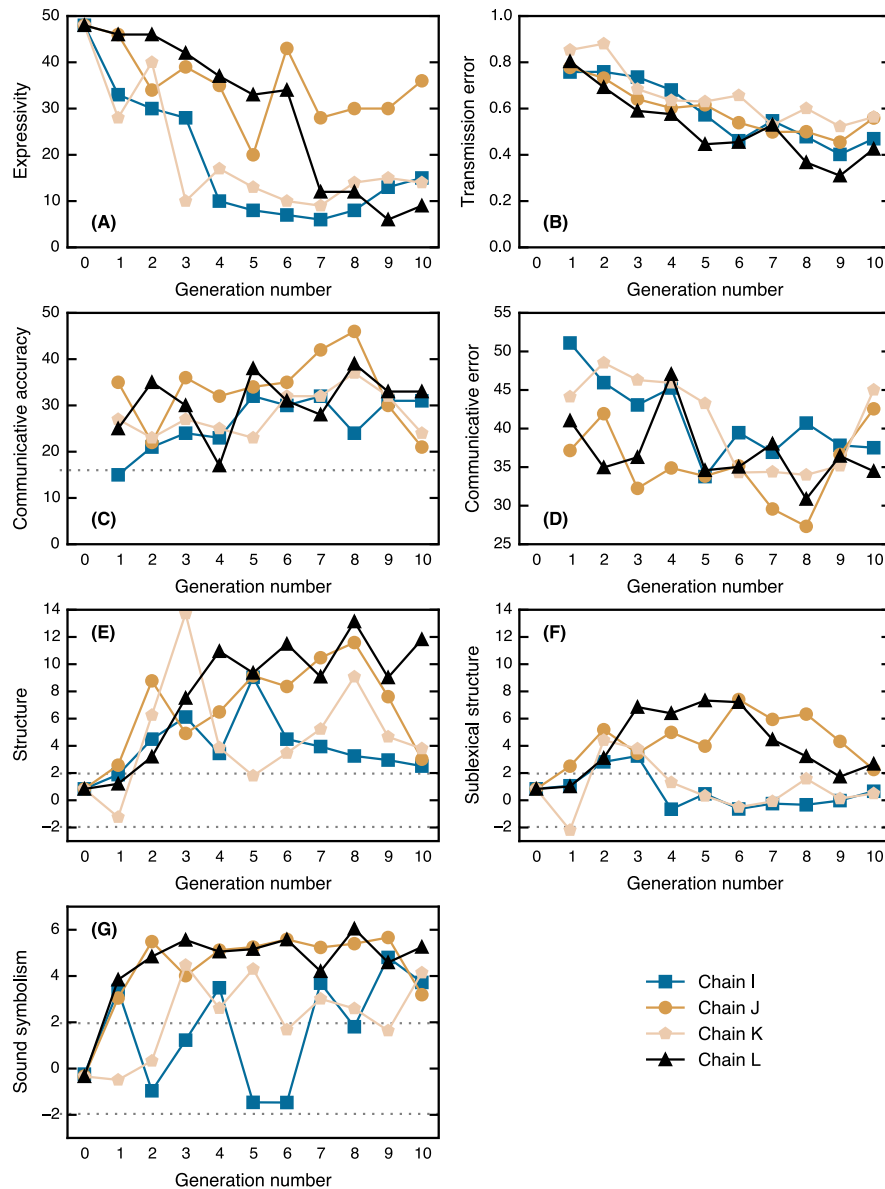


Fig. 8. Results of Experiment 3. (A) Expressivity: number of unique strings in the static set. (B) Levels of transmission error. (C) Number of correct trials (the dotted line indicates chance level). (D) Communicative error. (E) General structure. (F) Sublexical structure. (G) Shape-based sound symbolism. The dotted lines in (E), (F), and (G) give the upper and lower 95% significance levels; points lying outside of this interval are unlikely to be explained by chance.

pants got less than half of trials correct indicates that the task was particularly difficult and that there may be a ceiling on how well participants can perform, given the amount of training they receive and the length of time they communicate for. It is also likely that a pair of participants will not infer identical category boundaries, resulting in difficulty classifying nonprototypical members of a given category.

#### 4.2.4. Emergence of sublexical structure

The results for general structure are shown in Fig. 8E. Structure emerged very rapidly and remained high over the generations ( $L = 1,755$ ,  $m = 4$ ,  $n = 11$ ,  $p = .007$ ). Furthermore, Fig. 8F reveals that sublexical structure is present in Chains J and L, peaking at around Generation 6. To take one example, the language at Generation 6 in Chain L comprises five main units: *ba*, *da*, *fa*, *ma*, and *piku*. In nearly all cases, two or three of these units will be combined together to create a word. The way in which the words map onto the meaning space is shown in Fig. 9. Due to the large number of words, each Voronoi cell in the plot has been labeled to make the system easier to comprehend.

The pattern that immediately stands out is the tendency for labels represented by orange–yellow to cluster on the right-hand side of the plot. These triangles are labeled with words containing *piku* in initial and final position. There is also a clustering of reds and pinks corresponding to words containing *piku* in second or final position only. When *piku* occurs only once in the word, it usually indicates triangles that are small or somewhat thin (e.g., *bapiku*, *dapiku*, *fapiku*, *mapikuba*, *fadapiku*). When a word begins and ends with *piku*, it will usually refer to a very thin triangle with little area (e.g., *piku-fapiku*, *pikumapiku*, *pikumidpiku*). In fact, the three thinnest triangles are simply labeled *pikupiku*. These results suggest that reduplication, a common cross-linguistic phenomenon (Moravcsik, 1978), may play a role in intensifying meaning, perhaps through an iconic principle (double the *piku* corresponds to double the thinness; cf. Regier, 1998). Words with *da* in first position usually refer to triangles which are large and open (e.g., *dababa*, *dabafa*, *damafa*). However, when *da* occurs in second position, it often indicates that the triangle lies on the right-hand side of the bounding box (e.g., *fadaba*, *fadama*, *fadapiku*, *madada*, *madama*). Finally, words with *ma* in first position often correspond to triangles whose orienting spots point to the top-left corner of the bounding box (e.g., *madafa*, *mafaba*, *mamada*, *mapikufa*). However, these patterns are probabilistic; for each rule, exceptions can be identified.

Perhaps more interestingly, in many words, there appear to be meaningful subparts combined with nonmeaningful subparts. For example, the meanings of *fa* and *ma* in the words *pikufapiku* and *pikumapiku* are unclear. These subparts may be morphological residue like that found in cranberry morphs. Cranberry morphs are a class of morpheme that, for a given language, occur in only one word; as such, it is difficult to assign meaning to them without circular reference back to the word itself, calling into question the meaning of the term *morpheme* (traditionally, the smallest unit of meaning; see Aronoff, 1976, Chapter 2 for discussion of this issue). The classic example is the *cran* in the word *cranberry*, which has no independent meaning; instead it serves to distinguish cranberries from other types of berry. Similarly, the *fa* and *ma* in *pikufapiku* and *pikumapiku* may



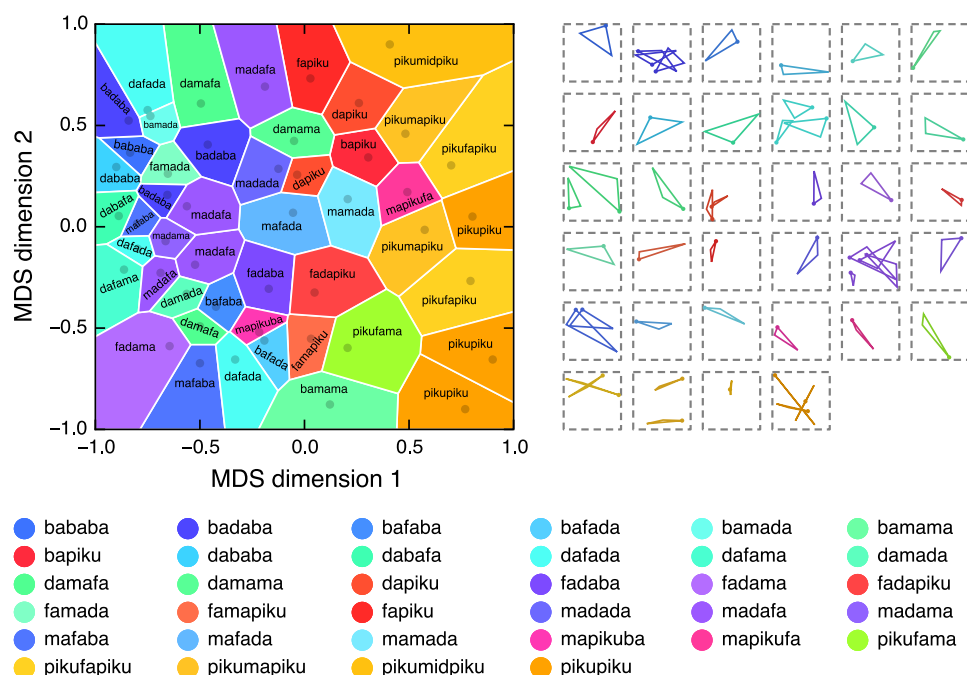


Fig. 9. Categorical structure of the meaning space at Generation 6 in Chain L. The plot on the left shows how the meaning space is discretized by the words in the language: Similarity in position represents similarity in meaning; similarity in color represents similarity in word form. On the right, all triangles in the static set are grouped by the word used to describe them.

express the idea, “I’m of the type *piku...piku*, but slightly different in a way I won’t explicitly specify.” For instance, the *fa* type of *piku...piku* is slightly longer and thinner than the *ma* type, but this correspondence does not appear to be productive across the language as a whole.

#### 4.2.5. Sound symbolism

Fig. 8G shows levels of shape-based sound symbolism, which are very strong and tend to emerge early in the chains. This is likely because the pair of participants can rely on a shared, implicit understanding of common sound-symbolic patterns to more accurately communicate with each other.

#### 4.2.6. Summary of Experiment 3

Introducing communication created a natural pressure for participants to be expressive. Expressivity remained higher than Experiment 1 and comparable to Experiment 2. Despite this, the learnability of the languages also remained high. Participants in at least two of the chains managed the pressures for expressivity and learnability by utilizing

string-internal structure that leverages the structure in the meaning space and sound-symbolic associations. Thus, in this experiment, where there was a pressure to maintain the diversity of signals due to the natural pressure from expressivity in addition to the pressure for learnability associated with transmission, sublexical structure emerged in addition to the general categorical structure observed in the previous experiments.

## 5. Discussion

In the Introduction, we claimed that our meaning space is a useful model of the natural world because the space of triangles is vast, continuous, and open-ended, properties that are present in objects that occur in the real world. For example, the vast set of items referred to by the English word *cup* forms a conceptual category that has fuzzy boundaries with neighboring concepts, such as *bowl*, *glass*, and *pitcher* (Labov, 1973). The dimensions of the conceptual space in which cups are represented may be either discrete (e.g., the presence or absence of a handle) or continuous (e.g., its size or shape). Similarly, our space of triangles potentially has both discrete (e.g., the quadrant in which the triangle is located) and continuous (e.g., the size or rotation of the triangle) dimensions with boundaries that are not well defined. Furthermore, our participants are unlikely to have strong preconceptions about how the space of triangles should be discretized. While geometrical terminology exists to describe the shape of triangles (equilateral, isosceles, and scalene) and their angles (acute, obtuse, and right-angled), these terms are not particularly useful in the context of our experimental paradigm, since they discretize the space of triangles based on artificial mathematical properties rather than naturally perceived features.

In Experiment 1, the languages that emerged discretized the meaning space into a small number of categories. Although the precise boundaries between categories varied from one chain to the next, the categories typically encoded the shape and size of the triangles; other features that could have been encoded—location or rotation in the plane—tended to be disregarded by the participants (see also Section 2 of Appendix S2 in the supplementary material). In fact, the naïve raters broadly responded to the space in the same way, rating the dissimilarity between triangles based on their shape and size properties (as evidenced by the dimensions of the MDS space). This is congruent with Landau, Smith, and Jones (1988), who showed that, when learning words, both children and adults are biased toward the shape of stimuli over their color, texture, or size. The process of collapsing categorical distinctions was taken to the extreme in one of the chains where a single word was used for all triangles by the final two generations. The process of collapsing categories is a valid strategy for maximizing compressibility (and therefore learnability), but the emergent languages in Experiment 1 were not expressive and would therefore be ill-suited to a world where one needed to reliably discriminate referents.

In Experiment 2, we placed a limit on the number of times a word could be reused, imposing an artificial expressivity pressure on the languages. This was intended to be equivalent to the pressure imposed in Kirby et al.'s (2008) second experiment. While the

number of unique strings remained high in Experiment 2, there was no evidence of the sublexical structure one would expect to find in a compositional system. In fact, the large amount of variation within each language even prevented stabilization on a set of categories in the meaning space. This result is strikingly different from the results reported by Kirby et al. (2008), who observed robust compositional structure under such a pressure. One explanation for this could be that, when the experimenter provides participants with a structured meaning space with unambiguous internal boundaries, single participants can simply transfer part of the meaning space structure onto the signals, cumulatively giving rise to compositional systems over generational time. In contrast, when participants are presented with an unstructured meaning space, as is the case here, the process of deriving structured signals becomes nontrivial. That being said, the artificial pressure used here is slightly different from that used by Kirby et al. (2008): The pressure involves direct instruction to participants—asking them to use different words when an arbitrary limit is reached—and does not maintain a one-to-one mapping between signal and meaning (a signal can map to up to three meanings in this experiment). The effects of such subtle differences are unclear and could be the subject of future work.

In Experiment 3, we added communication, which acts as a natural pressure for expressivity. In this experiment, each generation consisted of communicating participants who had the shared goal of maximizing their communicative accuracy. To achieve this, a language would be required that could encode a sufficient number of feature distinctions in order for the matching participant to correctly determine the target triangle. Like Experiment 2, expressivity remained high, but, unlike Experiment 2, the learnability of the languages also remained comparatively high and our measure of structure revealed that string-internal structure was present in two of the four chains. Thus, in this experiment, where there was a natural pressure to maintain a diverse set of signals, sublexical structure emerged in addition to the categorical structure observed in Experiment 1.

Nevertheless, it is difficult to describe the emergent sublexical structure as compositional, at least in terms of how compositionality is traditionally defined. A standard, theory-neutral definition of compositionality states that, “the meaning of a complex expression is determined by its structure and the meanings of its constituents” (Szabó, 2013). However, in our qualitative analysis of the emergent languages, it proved difficult to write simple grammars that could describe how to create composite strings with composite meanings because many of the mappings between form and meaning were highly probabilistic. In addition, in the exit questionnaire, many of our participants were unable to describe how the languages worked, suggesting instead that there were weak statistical tendencies in how form mapped onto meaning; one participant (Chain I, Generation 8, Subject A) remarked, “I think we had vague ideas of the template for each word, but we were pretty inconsistent.”

However, this is precisely how the lexicons of natural languages work. While polymorphemic words are compositional (either through inflection, *washed* = *wash* + *-ed*, or derivation, *happiness* = *happy* + *-ness*), monomorphemic words cannot be decomposed into smaller meaningful units. Furthermore, the extent to which polymorphemic words are compositional is also questionable. For example, Aronoff (1976, 2007) takes the view

that lexemes, even polymorphemic ones, are largely idiosyncratic. Sentences need to be highly compositional to provide language with its productivity, and the production of sentences is certainly a generative process, leading to combinations of words that have never been uttered before (although cf. Wray & Perkins, 2000). In contrast, the lexicon is stored in memory and many polymorphemic words have idiosyncratic meanings that have drifted from the sum of the parts from which they were originally derived. Aronoff therefore views polymorphemic lexemes as being only weakly compositional. While Aronoff's position may be a radical alternative to the classic view, it provides an alternative perspective on compositionality (or lack thereof) at the level of the lexeme.

The second linguistic property relevant to our results is de Saussure's (1959) *arbitrariness of the sign*, which states that the relationship between form and meaning is arbitrary and established only by convention among language users. In the context of language evolution, the importance of the arbitrariness of the sign was further solidified by Hockett (1960), who counted it among the design features of language. However, there are notable exceptions to this principle, which Cuskley and Kirby (2013) break down into conventional and sensory sound symbolism.<sup>7</sup>

Conventional sound symbolism refers to correspondences between signal and meaning that are set up by the historical relatedness of words. Such correspondences have been shown to contribute to the overall systematicity of natural languages using corpus-analytical techniques in both English (Monaghan, Shillcock, Christiansen, & Kirby, 2014) and Spanish (Tamariz, 2008). One example of this, which seems likely to contribute to such statistical correspondences, is phonesthesia—the phenomenon where monomorphemic words contain correspondences between sound and meaning. For example, English words beginning with *sn-* often have meanings relating to the nose (e.g., *sneeze*, *sniff*, *snore*, *snout*, etc.). Such words may possess shared etymologies that are obfuscated by the current state of the language and/or may be adopted precisely because of the correspondences they share with preexisting words in the lexicon. Bergen (2004) and Hutchins (1998) have shown in psycholinguistic experiments that the English phonesthemes have a psychological reality in the minds of native speakers, suggesting that they should be considered in a similar light to regular morphemes (see Kwon & Round, 2015, for some discussion).

The second type, sensory sound symbolism, involves correspondences between signal and meaning motivated by cross-modal or intramodal cognitive biases (see Lockwood & Dingemanse, 2015, for a review). This type of sound symbolism is particularly relevant to this study because it has been shown to facilitate word learning (e.g., Monaghan, Christiansen, & Fitneva, 2011; Nielsen & Rendall, 2012; Nygaard, Cook, & Namy, 2009; Parault & Schwanenflugel, 2006) and is frequently advanced as an explanation for the origin of language. We found significant levels of shape-based sound symbolism in the emergent languages. There was also some evidence for size-based sound symbolism in some of the languages using a conservative measure of size.

Compositionality and the arbitrariness of the sign are fundamental principles of language. However, recent research, briefly reviewed above, is suggestive of a more nuanced picture of language structure that our results are aligned with: Sound symbolic structure emerged in all three of our experiments, and, in Experiment 3, we found evidence of sub-

lexical structure that was not compositional in the traditional sense. In the early generations of Experiment 3, the pairs of participants shared little common ground, so they made use of iconic strategies, such as sound symbolism or reduplication. This gave rise to sublexical structure that peaked in each of the chains between Generation 2 and Generation 6. This sublexical structure then gradually started to drop away, perhaps—as Aronoff might argue—because the meanings of the words begin to drift from their compositional origins as “the sign gravitates to the word” (Aronoff, 1976, p. 14). That is to say, the words may be compositional early on and then start to lose this property as they begin to evolve idiosyncratic meanings not predictable from their component parts, just as in natural language where polymorphemic words cannot always be easily decomposed into smaller units of meaning.<sup>8</sup> We suggest that this aspect of compositionality, as well as a more complete understanding of how iterated learning builds morphemes out of noise—via an interim stage of statistical tendencies—is ripe for future exploration.

## 6. Conclusion

Our meaning space pushes the boundaries on the experimental study of iterated learning by avoiding several simplifications that previous experiments have made. Our meaning space is continuous, unstructured by the experimenter, vast in magnitude, and we do not prompt participants to make a certain number of categorical distinctions. Despite these features of the experimental setup, our first experiment showed that cultural evolution can deliver languages that categorize the meaning space under pressure from learnability. These languages had no string-internal structure but showed signs of containing sensory sound symbolic patterning. In our second experiment, and unlike previous studies, combining the pressure for learnability with an artificial pressure for expressivity did not lead to signals with internal structure. In our final experiment, we found that combining a pressure for learnability with a pressure for expressivity derived from a genuine communicative task gave rise to languages that use both categorization and string-internal structure to be both learnable and expressive. Unlike previous work, this emergent structure was sublexical rather than morphosyntactic, and as such bears similarities to certain aspects of natural lexicons, combining both conventional and sensory sound symbolism.

## 7. Acknowledgments

The authors thank Jennifer Culbertson, Amy Perfors, Bodo Winter, and two anonymous reviewers for their helpful comments on this work. The research also greatly benefited from numerous discussions with members of the Centre for Language Evolution at Edinburgh. JWC was funded by the Economic and Social Research Council (grant number ES/J500136/1) and a Carnegie–Cameron Taught Postgraduate Bursary. HC was funded by the British Academy for the Humanities and Social Sciences (grant number PDF110097).

## Notes

1. Although we do not test these predictions in this paper, we do use the notion of semantic convexity in our analyses. This notion states that “a subset  $C$  [i.e., a category] of a conceptual space  $S$  [i.e., a meaning space] is said to be *convex* if, for all points  $x$  and  $y$  in  $C$ , all points between  $x$  and  $y$  are also in  $C$ ” (Gärdenfors, 2000, p. 69). In other words, the members of a category form a single region of a meaning space in which it is possible to travel between any two members in a straight line without leaving the region.
2. The number of possible triangles in a finite space is uncountably infinite given the set of real numbers. However, the number of triangle stimuli in our meaning space is limited by the resolution of the display and ultimately by what participants are able to perceive as distinct. The latter is difficult to precisely quantify, but for the purpose of this paper, the space can be assumed to be vast in magnitude.
3. The minimum number of insertions, deletions, and substitutions that must be made to one string to transform it into another. The distance is normalized by dividing by the length of the longer string.
4. Correlation between the original dissimilarity ratings and the corresponding Euclidean distances in MDS space: .83. Stress-1 value: .25.
5. The denominator in Eq. 2 is the upper bound on the area of a triangle of given perimeter. When the ratio is 1, the triangle has maximum area given its perimeter and is therefore equilateral; as the ratio approaches 0, the triangle becomes increasingly thin and pointed.
6. Square root of the sum of squared distances from the centroid of the triangle to its vertices.
7. Cf. Dingemanse, Blasi, Lupyan, Christiansen, and Monaghan (2015), who refer to these notions under the terms “systematicity” and “iconicity.”
8. For example, the meaning of *reduce* is not predictable from *re-* and *-duce*, despite the fact that these morphemes appear in other English words: *receive*, *refer*, *repel*; *deduce*, *induce*, *produce* (Aronoff, 1976). However, the Latin etymology of these words indicates that they were indeed compositional in the past: *reducere* = *to lead back*, *referre* = *to carry back*, *repellere* = *to drive back*, etc.

## References

- Abbott, E. A. (1884). *Flatland: A romance of many dimensions*. London: Seeley.
- Ahlner, F., & Zlatev, J. (2010). Cross-modal iconicity: A cognitive semiotic approach to sound symbolism. *Sign Systems Studies*, 38, 298–348.
- Aronoff, M. (1976). *Word formation in generative grammar*. Cambridge, MA: MIT Press.
- Aronoff, M. (2007). In the beginning was the word. *Language*, 83, 803–830. doi:10.1353/lan.2008.0042
- Bergen, B. K. (2004). The psychological reality of phonaesthemes. *Language*, 80, 290–311. doi:10.1353/lan.2004.0056



- Bookstein, F. L. (1991). *Morphometric tools for landmark data: Geometry and biology*. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511573064
- Borg, I., & Groenen, P. J. F. (2005). *Modern multidimensional scaling: Theory and applications* (2nd ed.). New York, NY: Springer-Verlag. doi:10.1007/0-387-28981-X
- Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, 31, 489–558. doi:10.1017/S0140525X08004998
- Cuskley, C., & Kirby, S. (2013). Synesthesia, cross-modality, and language evolution. In J. Simner & E. M. Hubbard (Eds.), *The Oxford handbook of synesthesia* (pp. 869–899). Oxford, UK: Oxford University Press. doi:10.1093/oxfordhb/9780199603329.013.0043
- Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, iconicity, and systematicity in language. *Trends in Cognitive Sciences*, 19, 603–615. doi:10.1016/j.tics.2015.07.013
- Galantucci, B. (2005). An experimental study of the emergence of human communication systems. *Cognitive Science*, 29, 737–767. doi:10.1207/s15516709cog0000\_34
- Gärdenfors, P. (2000). *Conceptual spaces: The geometry of thought*. Cambridge, MA, MIT Press.
- Giordano, B. L., Guastavino, C., Murphy, E., Ogg, M., Smith, B. K., & McAdams, S. (2011). Comparison of methods for collecting and modeling dissimilarity data: Applications to complex sound stimuli. *Multivariate Behavioral Research*, 46, 779–811. doi:10.1080/00273171.2011.606748
- Hockett, C. F. (1960). The origin of speech. *Scientific American*, 203, 88–96.
- Horner, V., Whiten, A., Flynn, E., & de Waal, F. B. M. (2006). Faithful replication of foraging techniques along cultural transmission chains by chimpanzees and children. *Proceedings of the National Academy of Sciences of the USA*, 103, 13878–13883. doi:10.1073/pnas.0606015103
- Hurford, J. R. (1989). Biological evolution of the Saussurean sign as a component of the language acquisition device. *Lingua*, 77, 187–222. doi:10.1016/0024-3841(89)90015-6
- Hutchins, S. S. (1998). The psychological reality, variability, and compositionality of English phonesthemes (Doctoral dissertation). Available at ProQuest Dissertations and Theses database (UMI No. 9901857).
- Kay, P., Berlin, B., Maffi, L., Merrifield, W. R., & Cook, R. (2009). *The world color survey*. Stanford, CA: Center for the Study of Language and Information.
- Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, 336, 1049–1054. doi:10.1126/science.1218811
- Kirby, S. (2002). Learning, bottlenecks and the evolution of recursive syntax. In T. Briscoe (Ed.), *Linguistic evolution through language acquisition: Formal and computational models* (pp. 173–203). Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511486524.006
- Kirby, S. (2007). The evolution of meaning-space structure through iterated learning. In C. Lyon, C. L. Nehaniv, & A. Cangelosi (Eds.), *Emergence of communication and language* (pp. 253–267). London: Springer-Verlag. doi:10.1007/978-1-84628-779-4\_13
- 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 of the USA*, 105, 10681–10686. doi:10.1073/pnas.0707835105
- Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141, 87–102. doi:10.1016/j.cognition.2015.03.016
- Köhler, W. (1929). *Gestalt psychology*. New York: Liveright.
- Kovic, V., Plunkett, K., & Westermann, G. (2010). The shape of words in the brain. *Cognition*, 114, 19–28. doi:10.1016/j.cognition.2009.08.016
- Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30, 61–70. doi:10.1177/001316447003000105
- Kwon, N., & Round, E. R. (2015). Phonaesthemes in morphological theory. *Morphology*, 25, 1–27. doi:10.1007/s11525-014-9250-z
- Labov, W. (1973). The boundaries of words and their meanings. In C.-J. N. Bailey & R. W. Shuy (Eds.), *New ways of analyzing variation in English* (pp. 340–373). Washington, DC: Georgetown University Press.

- Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago, IL: The University of Chicago Press.
- Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, 3, 299–321. doi:10.1016/0885-2014(88)90014-7
- Laskowski, C. (2008). The emergence of a lexicon by prototype-categorising agents in a structured infinite world. In A. D. M. Smith, K. Smith, & R. Ferrer i Cancho (Eds.), *The evolution of language: Proceedings of the 7th international conference* (pp. 195–202). Singapore: World Scientific. doi:10.1142/9789812776129\_0025
- Lespinats, S., & Fertil, B. (2011). ColorPhylo: A color code to accurately display taxonomic classifications. *Evolutionary Bioinformatics*, 7, 257–270. doi:10.4137/EBO.S7565
- Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10, 707–710.
- Lockwood, G., & Dingemanse, M. (2015). Iconicity in the lab: A review of behavioral, developmental, and neuroimaging research into sound-symbolism. *Frontiers in Psychology*, 6, 1–14. doi:10.3389/fpsyg.2015.01246
- Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not just for talking. *Psychological Science*, 18, 1077–1083. doi:10.1111/j.1467-9280.2007.02028.x
- Malt, B. C., Sloman, S. A., & Gennari, S. P. (2003). Universality and language specificity in object naming. *Journal of Memory and Language*, 49, 20–42. doi:10.1016/S0749-596X(03)00021-4
- Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27, 209–220.
- Maurer, D., Pathman, T., & Mondloch, C. J. (2006). The shape of boubas: Sound-shape correspondences in toddlers and adults. *Developmental Science*, 9, 316–322. doi:10.1111/j.1467-7687.2006.00495.x
- 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, 325–347. doi:10.1037/a0022924
- Monaghan, P., Shillcock, R. C., Christiansen, M. H., & Kirby, S. (2014). How arbitrary is language? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369, 1–12. doi:10.1098/rstb.2013.0299
- Moravcsik, E. (1978). Reduplicative constructions. In J. H. Greenberg (Ed.), *Universals of human language: Word structure* (Vol. 3, pp. 297–334). Stanford, CA: Stanford University Press.
- Nielsen, A., & Rendall, D. (2012). The source and magnitude of sound-symbolic biases in processing artificial word material and their implications for language learning and transmission. *Language and Cognition*, 4, 115–125. doi:10.1515/langcog-2012-0007.
- Nuckolls, J. B. (1999). The case for sound symbolism. *Annual Review of Anthropology*, 28, 225–252. doi:10.1146/annurev.anthro.28.1.225
- Nygaard, L. C., Cook, A. E., & Namy, L. L. (2009). Sound to meaning correspondences facilitate word learning. *Cognition*, 112, 181–186. doi:10.1016/j.cognition.2009.04.001
- Page, E. (1963). Ordered hypotheses for multiple treatments: A significance test for linear ranks. *Journal of the American Statistical Association*, 58, 216–230. doi:10.1080/01621459.1963.10500843
- Parault, S., & Schwanenflugel, P. (2006). Sound-symbolism: A piece in the puzzle of word learning. *Journal of Psycholinguistic Research*, 35, 329–351. doi:10.1007/s10936-006-9018-7
- Pfers, A., & Navarro, D. J. (2014). Language evolution can be shaped by the structure of the world. *Cognitive Science*, 38, 775–793. doi:10.1111/cogs.12102
- Regier, T. (1998). Reduplication and the arbitrariness of the sign. In M. Gernsbacher & S. Derry (Eds.), *Proceedings of the 20th Annual Conference of the Cognitive Science Society* (pp. 887–892). Mahwah, NJ: Lawrence Erlbaum Associates.
- Rosch, E. H. (1973). Natural categories. *Cognitive Psychology*, 4, 328–350. doi:10.1016/0010-0285(73)90017-0
- de Saussure, F. (1959). *Course in general linguistics*. New York: Philosophical Library.



- Selten, R., & Warglien, M. (2007). The emergence of simple languages in an experimental coordination game. *Proceedings of the National Academy of Sciences of the USA*, 104, 7361–7366. doi:10.1073/pnas.0702077104
- Silvey, C. (2014). *The communicative emergence and cultural evolution of word meanings* (Unpublished doctoral dissertation). Edinburgh, UK: University of Edinburgh.
- Silvey, C., Kirby, S., & Smith, K. (2013). Communication leads to the emergence of sub-optimal category structures. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 1312–1317). Austin, TX: Cognitive Science Society.
- Smith, K. (2004). The evolution of vocabulary. *Journal of Theoretical Biology*, 228, 127–142. doi:10.1016/j.jtbi.2003.12.016
- Smith, K., & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition*, 116, 444–449. doi:10.1016/j.cognition.2010.06.004
- Szabó, Z. G. (2013). Compositionality. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2013 ed.). Available at: <http://plato.stanford.edu/entries/compositionality/>. Accessed September 27, 2015.
- Tamariz, M. (2008). Exploring systematicity between phonological and context-cooccurrence representations of the mental lexicon. *The Mental Lexicon*, 3, 259–278. doi:10.1075/ml.3.2.05tam
- Thompson, P. D., & Estes, Z. (2011). Sound symbolic naming of novel objects is a graded function. *The Quarterly Journal of Experimental Psychology*, 64, 2392–2404. doi:10.1080/17470218.2011.605898
- Verhoef, T. (2012). The origins of duality of patterning in artificial whistled languages. *Language and Cognition*, 4, 357–380. doi:10.1515/langcog-2012-0019
- Winters, J., Kirby, S., & Smith, K. (2015). Languages adapt to their contextual niche. *Language and Cognition*, 7, 415–449. doi:10.1017/langcog.2014.35
- Wray, A., & Perkins, M. (2000). The functions of formulaic language: An integrated model. *Language and Communication*, 20, 1–28. doi:10.1016/S0271-5309(99)00015-4
- Xu, J., Dowman, M., & Griffiths, T. L. (2013). Cultural transmission results in convergence towards colour term universals. *Proceedings of the Royal Society B: Biological Sciences*, 280, 1–8. doi:10.1098/rspb.2012.3073

### Supporting Information

Additional Supporting Information may be found online in the supporting information tab for this article:

**Appendix S1.** Experimental briefs

**Appendix S2.** Geometric measure of triangle dissimilarity

**Appendix S3.** MDS plots for all generations in all chains

### Appendix A: Online dissimilarity rating task

To measure the dissimilarity between pairs of triangles, we conducted an online experiment on the crowdsourcing platform CrowdFlower. A standard rating procedure was adopted, which is considered to be more reliable than other, more economical methods (Giordano et al., 2011). We collected dissimilarity ratings for the 1,128 pairs of triangles in the static set. The pairs of stimuli were randomly divided into 8 subsets of 141 pairs.

This was repeated 12 times, resulting in 96 subsets, each to be assigned to an individual participant. We paid a flat rate of \$0.50 for each of the 96 participants who completed the task. To access the task, participants had to correctly answer three simple entry questions, which evaluated their ability to understand basic English instructions; anyone who failed to answer these questions correctly was not allowed to enter the task. The participants were told that they would see pairs of triangles and would have to “rate how similar the two triangles are” using a slider control. The main part of the task was preceded by a 1 min familiarization stage in which participants were shown all 48 triangles in the static set to give them a sense of the maximum and minimum dissimilarity.

On each trial, the pair of triangles were presented side by side in 500×500-pixel dashed, gray bounding boxes. The slider control was located below the triangles and was labeled with *very similar* on one end and *very different* on the other; the direction of the scale was determined randomly for each participant. The slider had 1,001 levels of granularity, where 0 is maximally similar and 1,000 is maximally dissimilar. The participant could not proceed to the next trial until at least 3 s had passed and the slider control had been moved. After giving a rating, the participant had to press the enter key, which removed the triangles and slider from the screen, and then click a button labeled *next*, which was centered at the top of the screen; this forced the participant to move the mouse cursor to the top of the screen where it would be approximately equidistant from all points on the slider on the following trial.

There were six practice trials at the beginning of the experiment and three reliability trials randomly interspersed among the normal trials (for a total of 150 trials). In reliability trials, participants were shown identical triangles and should therefore have rated them with a low dissimilarity rating; this was included to monitor participants’ reliability. Due to a browser compatibility issue, a small portion of ratings (5.7%) were not recorded. After excluding these ratings, an average of 11.32 (*SD*: 1.48) independent ratings were collected for each pair of triangle stimuli. The median dissimilarity rating (on the 1,000-point scale) for reliability trials was 0, suggesting that participants were attending to the stimuli. Two participants were excluded because their mean ratings of reliability pairs were > 100.

The remaining 94 participants’ ratings were normalized in [0, 1] such that the ratings would use the entire width of the scale. The normalized ratings were then averaged together to produce a mean dissimilarity rating for each pair of triangles. Individual rater agreement was measured by correlating an individual participant’s ratings with the corresponding mean dissimilarity ratings for the 94 participants as a whole. Mean rater agreement was .7 (range: .22–.88). The three participants whose rater agreement was < .4 were then excluded, leaving a total of 91 participants.

The final distance matrix used in the main analysis was produced by averaging together the normalized ratings for the final 91 participants. There was an average of 10.72 (*SD*: 1.55) independent ratings per pair. Interrater reliability among the 91 participants was measured using Krippendorff’s alpha coefficient (Krippendorff, 1970), which is applicable where multiple raters each rate incomplete but overlapping subsets of the full data set. The value of this statistic was .41, which is quite low; however, this should not

be surprising given that participants were not instructed on specifically how to judge the dissimilarity between triangles, so some diversity in ratings was to be expected.

### **Appendix B: Dissimilarity judgments between target and selected triangles in Experiment 3**

Unless otherwise noted, this online experiment was identical to that described in Appendix A above. The 80 participants who took part in Experiment 3 selected the wrong triangle from the context array a total of 2,653 times. For a more granular measure of communicative error, we wanted to quantify the dissimilarity between the target and selected triangles in each of these cases. The 2,653 pairs were randomly divided into 21 subsets (14 subsets of 126 pairs and 7 subsets of 127 pairs). This was repeated 10 times, resulting in 210 subsets to be assigned to individual participants. We paid a flat rate of \$0.45 for each of the 184 participants who completed the task. There were six practice trials at the beginning and three reliability trials randomly interspersed among the normal trials (for a total of 135 or 136 trials).

The median number of independent ratings collected for each pair was 9 (range: 4–10). The median dissimilarity rating for reliability trials was 0. One participant was excluded because they rated all triangle pairs as having maximum dissimilarity. An additional 32 participants were excluded because their mean ratings of reliability pairs were > 100. The remaining 151 participants' ratings were normalized and averaged together to produce a mean dissimilarity rating for each pair of triangles. Mean rater agreement was .69 (range: .36–.87). The three participants whose rater agreement was < .4 were then excluded, leaving a total of 148 participants. The final dissimilarity ratings used in the main analysis were produced by averaging together the normalized ratings given by the final 148 participants. The mean number of independent ratings per pair of triangles was 7.04 (*SD*: 1.4). Krippendorff's alpha for interrater reliability was .37.

## 4.2 Summary of Paper 3

Paper 3 makes primarily two contributions. The first is its development of a fully continuous, open-ended meaning space. Other work in the literature has typically quantized the continuous space onto a grid (e.g. Canini et al., 2014; Perfors & Navarro, 2014; Silvey et al., 2015; J. Xu et al., 2013) because the iterated learning paradigm demands a design in which system-wide measurements are taken on a consistent set of stimuli. But quantization has its limitations. Recall that the Shepard circles in Paper 2 were quantized onto an 8×8 grid; it seems plausible that, after 160 training trials, many of the participants will have realized that they were only being exposed to just eight particular size and angle distinctions, which may prompt them to use, for example, rule-based, as opposed to say prototype-based, learning strategies. In Paper 3 we get around this issue by fixing a ‘static set’ of triangles which every participant labels and which may be used for measuring the variables of interest to us, while allowing the languages to evolve through the constantly changing ‘dynamic sets’ (see Fig. 2 on page 116). In addition, the complex structure of the space is such that there are no preestablished semantic categories, obvious dimensions, or natural kinds – these emerge through the process of iterated learning (see Lupyan, 2017, for recent experimental work on the human conceptualization of triangles).

The second contribution of the paper is its demonstration that genuine communicative interaction is required for the emergence of higher-level compositional structure. The paper can therefore be viewed as a replication of Kirby et al. (2015) but with the additional feature that the underlying semantic categories emerge alongside the compositional structure. While previous work in the iterated learning literature has treated categorical and compositional structure separately, Paper 3 is the first – to our knowledge – to show how the two arise together. In addition, the paper shows how artificial languages can adapt in unexpected ways; in designing the experiments, we had not anticipated, for example, that sound symbolic structure might emerge as a partial solution to the problem of making languages more learnable.<sup>17</sup>

In the remainder of this chapter, I discuss aspects of the project not discussed in the paper. First, in Section 4.3 I discuss the structure measures in more detail and

<sup>17</sup> Indeed, my later experiments reported in Paper 2 were designed to prevent the emergence of sound symbolic structure by randomizing the signal–category mapping at every generation (see page 70).

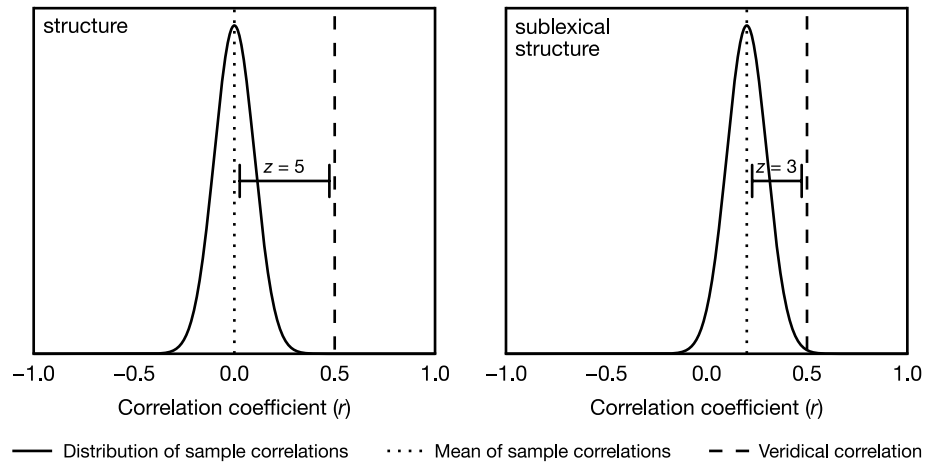
demonstrate in an artificial test-case that they are indeed able to pick up on emergent categorical and compositional structure. Section 4.4 reports some errata in relation to Page's test and provides recomputed statistics following the same methods used in Paper 2. In Section 4.5, I reanalyse the data in terms of simplicity and informativeness and show that the pressure from learning leads to simpler, less informative languages in comparison to the pressure from communication. Section 4.6 concludes the chapter.

### 4.3 Structure and the Mantel Test

The method used to measure structure in Paper 3 (page 119) has become the *de facto* standard in the iterated learning literature. Shillcock, Kirby, McDonald, and Brew (2001) first used the method to demonstrate that the phonological distances between monosyllabic English words are correlated with the semantic distances between those words; words that are phonologically similar tend to have similar meanings (see also Monaghan, Shillcock, Christiansen, & Kirby, 2014; Tamariz, 2008). In the iterated learning literature, the method was first used in early computational simulations (e.g. Brighton, Smith, & Kirby, 2005) and later in the experimental analogues of those simulations (e.g. Kirby et al., 2008, 2015). However, the method does in fact have an earlier provenance in the form of the Mantel test (Mantel, 1967), which is used to calculate the significance of the correlation between two distance matrices.<sup>18</sup>

Our measure of structure correlates the pairwise distances between words with the corresponding pairwise distances between meanings. An illustration of this is provided in Fig. 4.1. In this hypothetical example, the veridical Pearson correlation between the meaning distances and string distances is 0.5 (dashed line). The mapping between meanings and strings is then randomized, and the correlation is remeasured; this is then repeated some large number of times (100,000 in the paper), yielding a distribution of correlation coefficients centred around zero, as shown by the solid curve in the graph on the left. The distribution is normally distributed around zero because, under random string-meaning mappings, there is, on average, no correlation between meaning similarity and string similarity. The level of structure in the language is then estimated by

18 The standard approach to obtaining a *p*-value for a correlation coefficient is not suitable in the case of correlating two distance matrices because the correlation of two such matrices violates the assumption that each datapoint is independent of the others; moving a single point in some underlying space affects not one but many of the distances in the distance matrix (see Cornish, 2011, p. 91–94).



**Figure 4.1:** Examples of the regular structure measure (left) and the sublexical structure measure (right) applied to the same hypothetical dataset. Dashed lines show the veridical correlation between string distances and meaning distances ( $r = 0.5$ ). The correlation under a sample of random string–meaning mappings is shown by the solid curve on the left; when the mapping between signal and meaning is randomized there is, on average, no correlation ( $r \approx 0.0$ ). Under the sublexical structure measure, the mapping is randomized between category labels and categories, such that the sample languages retain any categorical structure, resulting in a higher correlation ( $r \approx 0.2$ ). If the veridical is significantly greater than the sample, the language must contain general or sublexical structure.

how far the veridical correlation is from the mean of the Monte-Carlo sample correlations, computed as a  $z$ -score. Since the veridical correlation is significantly greater than the distribution of the Monte-Carlo permutations (here,  $z = 5$ ), there must be systematic structure present in the language; words with similar form have similar meaning.

However, this measure is not able to disambiguate between categorical structure in the meaning space and structure internal to the strings themselves. Paper 3 introduces a novel method for distinguishing between these two forms of structure, which is illustrated on the right-hand side of Fig. 4.1. Under this method, rather than randomize the mapping between all 48 meanings and their corresponding signals, the mapping is randomized between category labels (i.e. the unique set of words in the language) and their corresponding categories (i.e. sets of meanings). In other words, the triangles that were grouped together into a particular category remain together but are assigned one of the strings at random. As such, the randomized mappings in the Monte-Carlo sample retain any categorical structure, yielding a distribution of correlation coefficients with a mean that is greater than zero (0.2 in the example). However, since the veridical correlation is still significantly greater than the distribution of Monte-Carlo permuta-

tions ( $z = 3$ ), there must be additional structure present in the strings themselves. We refer to this additional structure as sublexical structure.

To demonstrate that this measure works, I first coded each of the 48 triangles in the static set as being either above average (0) or below average (1) in terms of location, rotation, size, and shape (using the measures described in Appendix E), which resulted in 15 ‘natural kinds’ (triangles that are similar in terms of all four features) as shown in Table 4.1. I then generated two languages: a categorical language in which each of the 15 natural kinds is labelled by a unique string, and a compositional language in which each of the feature values is represented by a particular letter. For example, triangles that are bigger than average have the letter *t* in third position, and triangles that are smaller than average have the letter *p* in third position. The two languages use the exact same set of words, but under the compositional language, the words are concatenated from a set of units that have a systematic relationship with particular features of the triangles.

Using the geometric measure of triangle dissimilarity described in Appendix E to provide the semantic distances, the categorical language has a structure score of 7.54 (there is a significant level of structure) but a sublexical structure score of 1.11 (there is no significant level of sublexical structure); in contrast, the compositional language has a structure score of 20.81 (there is a significant level of structure) and a sublexical structure score of 7.01 (there is also a significant level of sublexical structure). This demonstrates that the measures can, in principle, detect the kinds of structure we would expect to find.

#### 4.4 Problems with Page’s Test: Some Errata

Page’s test is a nonparametric test of monotonically ordered differences in rank (Page, 1963). Until recently, Page’s test was commonly used in iterated learning studies to test for cumulativity – an increasing (or decreasing) trend over generational time (e.g. Caldwell & Millen, 2008; Kirby et al., 2015; Smith & Wonnacott, 2010; Verhoef, 2012). However, subsequent to the publication of Paper 3, Stadler (2017) pointed out that the use of Page’s test in iterated learning studies is highly problematic because Page’s test is not – and was never designed to be – a test of trend, and a single increase (or decrease) from one generation to the next may be sufficient to reject the null hypothesis. This is especially problematic given that, in typical iterated learning experiments, the ‘zeroth

**Table 4.1:** Binary feature values of the 48 triangles in the static set, sorted into ‘natural kinds’, and then labelled by a categorical or compositional language

Triangle	Location	Rotation	Size	Shape	Kind	Cat. language	Comp. language
5	0	0	0	0	1	<i>napi</i>	<i>kati</i>
29	0	0	0	0	1	<i>napi</i>	<i>kati</i>
12	0	0	0	0	1	<i>napi</i>	<i>kati</i>
11	0	0	0	0	1	<i>napi</i>	<i>kati</i>
37	0	0	0	0	1	<i>napi</i>	<i>kati</i>
34	0	0	0	0	1	<i>napi</i>	<i>kati</i>
16	0	0	0	0	1	<i>napi</i>	<i>kati</i>
2	0	0	0	1	2	<i>kopu</i>	<i>katu</i>
7	0	0	0	1	2	<i>kopu</i>	<i>katu</i>
48	0	0	1	0	3	<i>nati</i>	<i>kapi</i>
33	0	0	1	1	4	<i>nopu</i>	<i>kapu</i>
8	0	0	1	1	4	<i>nopu</i>	<i>kapu</i>
23	0	0	1	1	4	<i>nopu</i>	<i>kapu</i>
21	0	0	1	1	4	<i>nopu</i>	<i>kapu</i>
27	0	1	0	0	5	<i>napu</i>	<i>koti</i>
6	0	1	0	0	5	<i>napu</i>	<i>koti</i>
26	0	1	0	0	5	<i>napu</i>	<i>koti</i>
36	0	1	1	0	6	<i>notu</i>	<i>kopi</i>
30	0	1	1	0	6	<i>notu</i>	<i>kopi</i>
9	0	1	1	1	7	<i>natu</i>	<i>kopu</i>
19	0	1	1	1	7	<i>natu</i>	<i>kopu</i>
25	1	0	0	0	8	<i>koti</i>	<i>nati</i>
43	1	0	0	0	8	<i>koti</i>	<i>nati</i>
32	1	0	0	1	9	<i>noti</i>	<i>natu</i>
42	1	0	0	1	9	<i>noti</i>	<i>natu</i>
4	1	0	1	0	10	<i>kapi</i>	<i>napi</i>
38	1	0	1	1	11	<i>nopi</i>	<i>napu</i>
31	1	0	1	1	11	<i>nopi</i>	<i>napu</i>
13	1	0	1	1	11	<i>nopi</i>	<i>napu</i>
17	1	0	1	1	11	<i>nopi</i>	<i>napu</i>
41	1	1	0	0	12	<i>kapu</i>	<i>noti</i>
46	1	1	0	0	12	<i>kapu</i>	<i>noti</i>
20	1	1	0	0	12	<i>kapu</i>	<i>noti</i>
3	1	1	0	0	12	<i>kapu</i>	<i>noti</i>
22	1	1	0	1	13	<i>katu</i>	<i>notu</i>
1	1	1	0	1	13	<i>katu</i>	<i>notu</i>
39	1	1	0	1	13	<i>katu</i>	<i>notu</i>
24	1	1	0	1	13	<i>katu</i>	<i>notu</i>
14	1	1	0	1	13	<i>katu</i>	<i>notu</i>
45	1	1	1	0	14	<i>kati</i>	<i>nopi</i>
47	1	1	1	0	14	<i>kati</i>	<i>nopi</i>
15	1	1	1	0	14	<i>kati</i>	<i>nopi</i>
28	1	1	1	0	14	<i>kati</i>	<i>nopi</i>
10	1	1	1	0	14	<i>kati</i>	<i>nopi</i>
44	1	1	1	0	14	<i>kati</i>	<i>nopi</i>
18	1	1	1	1	15	<i>kopi</i>	<i>nopu</i>
35	1	1	1	1	15	<i>kopi</i>	<i>nopu</i>
40	1	1	1	1	15	<i>kopi</i>	<i>nopu</i>



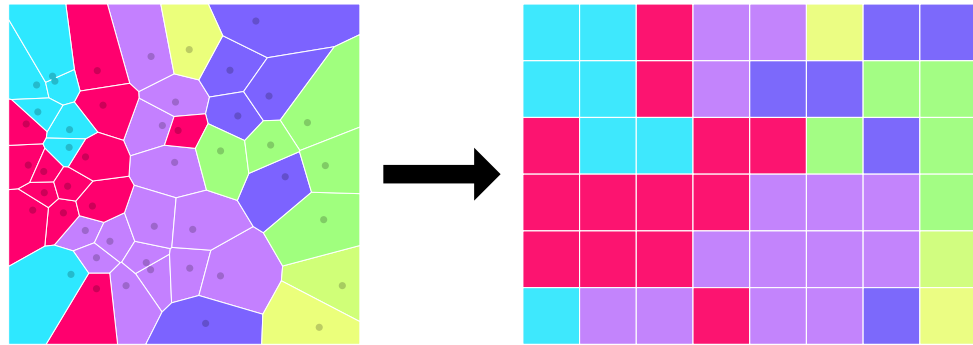
**Table 4.2:** Page's test statistics reported in Paper 3 and their LMER equivalents

Page	Variable	Page's test result	LMER model comparison result
117	Exp. 1, expressivity	$L = 1993, p < .001$	$\beta = -3.92 \pm 0.58, \chi^2 = 11.04, p < .001$
119	Exp. 1, trans. error	$L = 1514, p < .001$	$\beta = -0.066 \pm 0.01, \chi^2 = 8.86, p = .003$
119	Exp. 1, structure	$L = 1472, p < .001$	$\beta = 0.75 \pm 0.32, \chi^2 = 4.07, p = .044$
124	Exp. 2, trans. error	$L = 1415, p < .001$	$\beta = -0.01 \pm 0.00, \chi^2 = 8.34, p = .004$
128	Exp. 3, trans. error	$L = 1503, p < .001$	$\beta = -0.04 \pm 0.01, \chi^2 = 11.8, p < .001$
128	Exp. 3, comm. acc.	$L = 1321.5, p = .021$	<b><math>\beta = 0.75 \pm 0.39, \chi^2 = 3.2, p = .074</math></b>
128	Exp. 3, comm. error	$L = 1356, p = .004$	$\beta = -0.8 \pm 0.33, \chi^2 = 4.28, p = .039$
130	Exp. 3, sub. struct.	$L = 1755, p = .007$	<b><math>\beta = -0.1 \pm 0.11, \chi^2 = 0.8, p = .371</math></b>

generation' is often markedly different from the data obtained from actual participants, making a single generational change in the hypothesized direction very likely. By simulating random datasets in which there are four chains and 11 generations (as is the case in Paper 3), such that generation 0 is always lower than generation 1, we find that a significant result at  $\alpha = .05$  is obtained by chance about 47% of the time!

I have rerun all eight statistics reported in the paper using the same method described in Paper 2 (see page 74), which Winter and Wieling (2016) recommend for iterated learning experiments. Specifically, a linear mixed-effects regression analysis was used to test for an effect of generation on a particular variable of interest with chain as a random effect and by-chain random slopes for the effect of generation. *P*-values were obtained by likelihood ratio tests of the full model against a null model without the effect in question. These statistical results are given in Table 4.2, alongside the original Page's test statistics reported in the paper and the page numbers for reference.

Of these new statistical results, two were not significant (those highlighted in bold in Table 4.2). Firstly, communicative accuracy in Experiment 3 was *not* found to increase with generation. This revised finding is not especially important to the interpretation of the paper because we also provided a more fine-grained measure of communicative accuracy based on the dissimilarity between the director's target triangle and the matcher's selected triangle, and under this more precise measure, there *is* a significant decrease in communicative error, as shown in the table. Secondly, sublexical structure in Experiment 3 was *not* found to increase with generation, which on inspection of the plot on page 129 is not especially surprising given that sublexical structure only really emerged in two of the four chains, and even then it was gradually eroded after generation 6. Nevertheless, this is not fatal to the overall interpretation of the paper because it



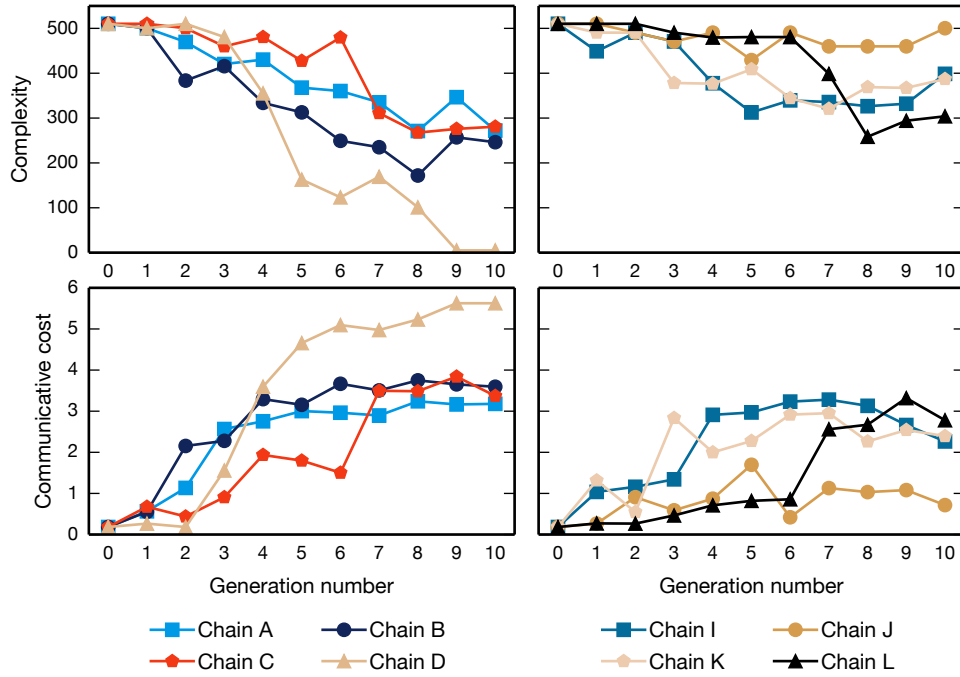
**Figure 4.2:** For the purpose of measuring complexity using the rectangle code, the languages from Paper 3 may be transformed from the continuous space to a discrete 8×6 grid, which approximates the structure of the meaning space. The colour coding in this example is for Generation 10 in Chain A.

only suggests that there was no *cumulative* effect on sublexical structure; the more important point, however, is that sublexical structure did emerge in Experiment 3, while it did not in Experiment 1: Iterated learning alone gives rise to categorical structure, while the addition of a communicative task can give rise to sublexical structure.

## 4.5 Simplicity and Informativeness

At the end of Chapter 2, I suggested that research on the evolution of semantic category systems can benefit from an approach that unifies the measures used by Kirby and colleagues and Regier and colleagues (see Fig. 2.15). As we have seen, Paper 3 included a communicative task in Experiment 3, so we would expect to find that, under pressures from both induction and interaction, the languages will gravitate towards the optimal frontier. Although the experiments in Paper 3 were not designed with measuring complexity and communicative cost in mind, it is nevertheless interesting to see if this prediction is borne out in the results. To perform these analyses, the 48 meanings in the static set were mapped onto a discrete 8×6 grid which approximates the continuous MDS solution, making it possible to approximate the complexity of the languages in the rectangle code. For example, Fig. 4.2 illustrates the language from Generation 10 in Chain A (depicted in the paper on page 120) and its mapping onto the discrete grid.

The results are shown in Fig. 4.3. Complexity decreases with generation in Experiment 1 (transmission-only;  $\beta = -36.74 \pm 9.05$ ,  $\chi^2 = 7.48$ ,  $p = .006$ ) and Experiment 3

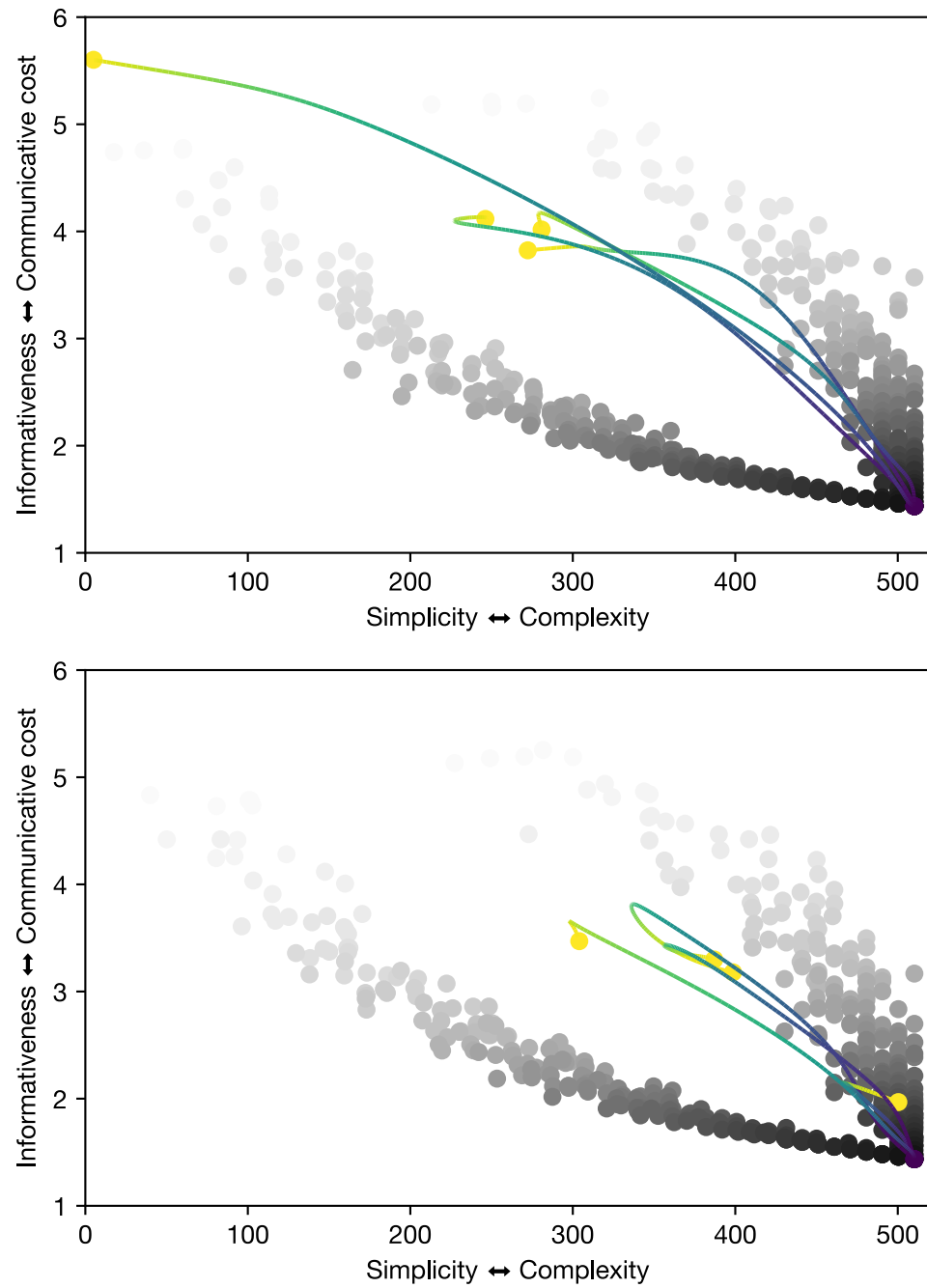


**Figure 4.3:** Complexity and communicative cost in Experiment 1 (transmission-only; left) and Experiment 3 (transmission with communication; right).

(transmission with communication;  $\beta = -14.57 \pm 5.47$ ,  $\chi^2 = 4.85$ ,  $p = .028$ ). Over generations, the category systems become simpler, although this is to be expected given that the category systems start out maximally complex (48 categories). However, the languages in Experiment 1 become more simple, as predicted by the fact that there is no communicative component, while the languages in Experiment 3 remain a little more complex. Communicative cost (bottom of Fig. 4.3) increases with generation in Experiment 1 ( $\beta = 0.3 \pm 0.07$ ,  $\chi^2 = 7.67$ ,  $p = .006$ ) and Experiment 3 ( $\beta = 0.15 \pm 0.05$ ,  $\chi^2 = 5.06$ ,  $p = .025$ ); again, this is to be expected given that the languages start out maximally informative (48 categories), but the increase is more pronounced in Experiment 1, while the languages in Experiment 3 remain somewhat more informative (especially Chains J and L).

## 4.6 Conclusion to Chapter 4

This chapter provides experimental evidence to show that communicative interaction is the driving force behind informative languages structure. Fig. 4.3 shows that iterated



**Figure 4.4:** Evolutionary trajectories through simplicity–informativeness space for Experiment 1 (top) and Experiment 3 (bottom). Each chain transitions from a purple dot (a random 48-category language) to a yellow dot over ten generations, and the curves show smoothed trajectories through the space. The grey dots represent randomly generated systems that are convex (left) or random (right) with varying numbers of categories (darker grey = more categories). Together, the grey dots approximately delimit the space of possible languages. All chains become simpler and less informative, but less so in Experiment 3, which includes an interactional component.

learning alone tends towards greater simplicity – indeed, one of the chains reached the trivial partition after nine generations becoming maximally costly to use. In comparison, when the pressure from interaction is included, the languages remain more complex, but also more informative. Informativeness is maintained by retaining as many unique signals as possible, but these signals develop a statistical form of compositional structure, which is more compressible, in response to the learning pressure that is also present.

As argued in this thesis (see Fig. 2.14 on page 47), the combination of both pressures should amount to a movement towards the optimal frontier as the languages find an optimal balance between simplicity on the one hand and informativeness on the other. Fig. 4.4 shows the evolutionary trajectories that the iterated learning chains take in simplicity–informativeness space in Experiment 1 (transmission-only; top) and Experiment 3 (transmission with communication; bottom). Of course, since the initial randomly-generated languages were maximally complex and maximally informative (every meaning had its own randomly generated category label), looking at the data in this way is somewhat limited. Nevertheless, it is possible to see that the languages have a less dramatic westwards expansion when the communicative pressure is present.



## Chapter 5

# Conclusion

Seldom do more than a few of nature's secrets give way at one time.

— Claude Shannon (1956)

To a certain extent I have done in this thesis precisely what Shannon told us not to do: Jump on the information theory bandwagon. In Chapter 2, we saw how the two principal pressures that shape language structure, induction and interaction – or equally the two halves of the simplicity–informativeness tradeoff – may be rendered in information-theoretic terms:—

First, while Bayes' theorem tells us that a rational learner ought to weigh up the likelihood and prior, it is information theory – or rather its progeny, algorithmic complexity theory and the MDL principle – that offers insight into how that prior ought to be set. A learner who has no particular expectations in some domain – for example, learning that this kind of circle is a *zix* and that kind of circle is a *zox* – should nevertheless place greater weight on simple explanations because, ultimately, 'nature does nothing in vain', so simple explanations are inherently more probable (Chater & Vitányi, 2007; Culbertson & Kirby, 2016; Li & Vitányi, 2008; Rissanen, 1978; Solomonoff, 1964a).

And second, in terms of communicative interaction, information theory has something to contribute once again. Regier and colleagues' communicative cost framework formalizes the precision with which a language or category system is able to convey meaning, capturing expected information loss during the transmission of meaning from one mind to another through the lossy medium of language. Communicative cost can

be seen as quantifying what Shannon and Weaver (1949, p. 4) called ‘the semantic problem’: ‘How precisely do the transmitted symbols convey the desired meaning?’ Moreover, Regier and colleagues’ body of work has shown that language is optimized not just in terms of informativeness (communicative cost) but also in terms of simplicity (complexity) in a wide variety of domains: kinship (Kemp & Regier, 2012), spatial relationships (Khetarpal et al., 2013), numeral systems (Y. Xu & Regier, 2014), colour terms (Regier et al., 2015), and container names (Y. Xu et al., 2016).

So perhaps, then, Shannon was overly pessimistic when he said that information theory cannot be expected to shed light on more than a few of nature’s secrets; it certainly seems to me that the language and cognitive sciences have benefited greatly from its simple, elegant formalism and predictive power over the past 70 years.

## 5.1 Recapitulation

Chapter 1 introduced some of the key ideas behind this thesis. Firstly, although language is manifestly underpinned by the unique configuration of human biology, languages are also socially learned, used, and transmitted, and their structure develops, at least in part, in response to these processes. Moreover, the structures that languages adopt through cultural evolution modify the biological fitness landscape, resulting in coevolutionary dynamics (see e.g. Christiansen & Chater, 2008; de Boer & Thompson, 2018; Richerson & Boyd, 2005). Ascertaining how these things fit together is one of the core problems that the language sciences are engaged in, and the argument put forward in this thesis is that the human learning mechanism has a tendency to simplify language and iron out its irregularities, while the need for precise communication (or indeed precise internal representation) prevents that process of simplification from getting out of hand. Many of the structural properties we recognize in natural languages and conceptual systems emerge from this neverending tug of war. The iterated learning framework has provided a successful paradigm in which these kinds of issues can be explored, particularly in terms of the contribution made by human inductive reasoning. More recent work in this literature has combined iterated learning with communication games, providing a simple but powerful model of the core factors shaping language.

Chapter 2 took these evolutionary ideas as a starting point and asked, how do inductive reasoning and communicative interaction shape the way humans partition the



world into discrete categories? The chapter attempted an answer to this question in four sections. Section 2.1 outlined what we mean by concepts and categories and described some of the key ideas in the literature about what makes a good category system (see also Douven & Gärdenfors, in press, for more thoughts on this). Section 2.2 formally established the link between learning and simplicity (measured in terms of complexity or compression) and showed that simple category systems are comprised of few categories with compact structure. Section 2.3 then formally established the link between communication and informativeness (measured as communicative cost) and showed that informative category systems are comprised of many categories with compact structure. Finally, in Section 2.4, we brought this all together under the simplicity–informativeness tradeoff: Semantic category systems evolve to find an optimal balance primarily in terms of expressivity – the number of categories the system contains – but the compactness property can only ever increase since it is favoured by both pressures.

Chapter 3 applied these theoretical ideas to a concrete issue that arose in the literature – namely, the surprising result reported by Carstensen et al. (2015). The authors contend, contrary to prior work, that learning favours informativeness. Ultimately, we argue that the authors have come to this conclusion by attributing to informativeness something that is more parsimoniously attributed to simplicity. Aside from this, however, Chapter 3 made several other contributions. We saw how finding a compressed encoding of a system of concepts accurately predicts (a) how easily that system will be learned and (b) the kinds of system that will emerge from iterated learning over generational time, replicating a slew of findings from both the category learning and iterated learning literatures. For example, we find that one-dimensional concepts are easier to learn than two-dimensional ones; we find that intergenerational information loss – brought about through the bottleneck, lack of exposures, or noise – contributes to greater convergence on the prior bias; and we find that the human inductive bias is best characterized by a preference for simplicity. Each on its own is not a new finding, but as a whole this project provides a foundation on which future work on the evolution of language and semantic category systems can build.

Finally, Chapter 4 showed how communicative interaction constrains iterated learning, preventing the process of degeneration – convergence to an inductive bias for simplicity – from getting out of hand; when participants must use a language to accomplish

some goal, they impart a pressure for informativeness. The first experiment showed that it is learning that delivers the basic hallmark properties of conceptual structure, category sparsity and compactness, while the third experiment showed that it is communication that delivers expressivity by means of higher-level forms of structure that optimize for the simplicity–informativeness tradeoff. This higher-level form of structure is compositionality, which is both simple and informative, although the form of compositional structure we identified showed only statistical tendencies towards a complete compositional system. In the process of demonstrating this, the project also highlighted that the iterated learning paradigm can be extended to more realistic meaning spaces in which the semantic categories are not provided for free by the experimenter, thereby ruling out one potential criticism of the paradigm.

## 5.2 Future directions

There are primarily two directions I would like to take the work presented in this thesis. The first direction would seek to better understand the human inductive bias, and the effects it has on cultural phenomena, by direct comparison to another primate species. Little has been said in this thesis about animal studies, but clearly there is something in the human biological endowment that sets us apart in terms of language, and the only way to elucidate what that might be is through comparison with other species. One promising line of enquiry comes from work by Claidière et al. (2014), who have successfully applied the iterated learning paradigm to a nonhuman primate species, the Guinea baboon. In this study, baboons completed a pattern reproduction task on touch screen computers, with the production output of one animal becoming the training input to the next animal in an iterated learning chain. Using this procedure, Claidière et al. (2014) provided the first evidence that a nonhuman primate species can exhibit three fundamental aspects of cultural evolution: a progressive increase in performance, the emergence of systematic structure, and the presence of lineage specificity.

However, the study was concerned with the general effect of iterated learning and did not specifically study structural properties relating to language. Understanding the extent to which the basic structural properties of language can be explained by cultural evolutionary dynamics would tell us whether the emergence of language is dependent on possessing the right cognitive endowment or possessing the right socio-cultural

mechanisms required in use and transmission. For example, direct replication of the iterated learning experiment in Paper 2 in a nonhuman primate would inform us that culturally defined categories can emerge in another species provided that the animals are given the right cultural scaffolding.

The second direction I would like to take the work presented herein would be to further refine the model introduced in Paper 2. This could include relatively small refinements, such as generalizing the framework to account for both separable and integral meaning dimensions, or larger scale development, such as integrating a model of communicative interaction. Currently, the iterated learning model I presented only includes learning and does not make predictions about what would happen when agents must communicate. This could add interesting dynamics to the model because we should not forget that learning usually takes place in the context of communication; children learn language while engaged in communicative interaction with caregivers, not through passive exposure to meaning–signal pairs. In other words, the agents would learn not just that some signal maps to some meaning but also about how effective signals are in eliciting certain responses, providing a doorway for informativeness. Work by Frank and Goodman (2012, 2014) might provide a useful starting point in modelling this aspect of language in a agent-based framework. However, it might also be possible to take a deeper, more mathematical approach to modelling the simplicity–informativeness tradeoff, as recently suggested by Zaslavsky, Kemp, Regier, and Tishby (2018), who formalize the tradeoff in terms of the information bottleneck principle (Shamir, Sabato, & Tishby, 2010; Tishby, Pereira, & Bialek, 1999), an information-theoretic approach to finding the best tradeoff between accuracy and complexity.

### 5.3 Final thoughts

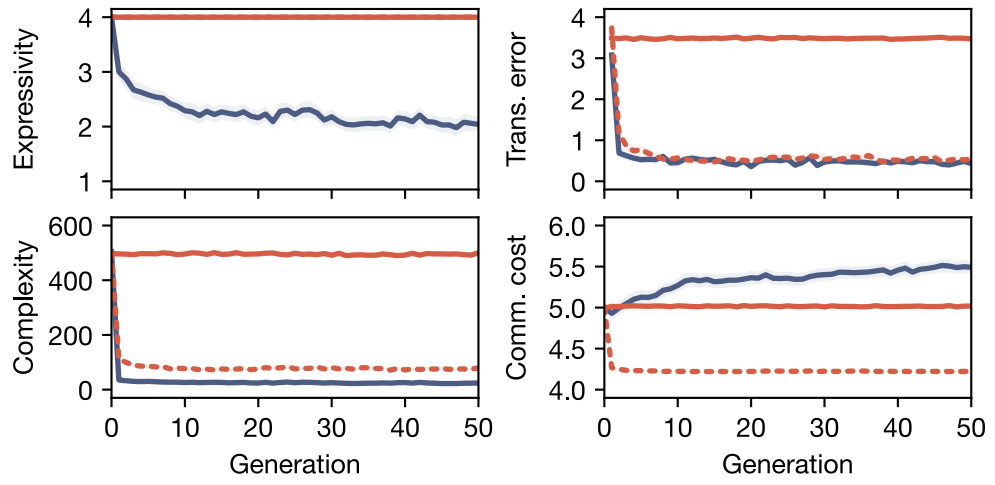
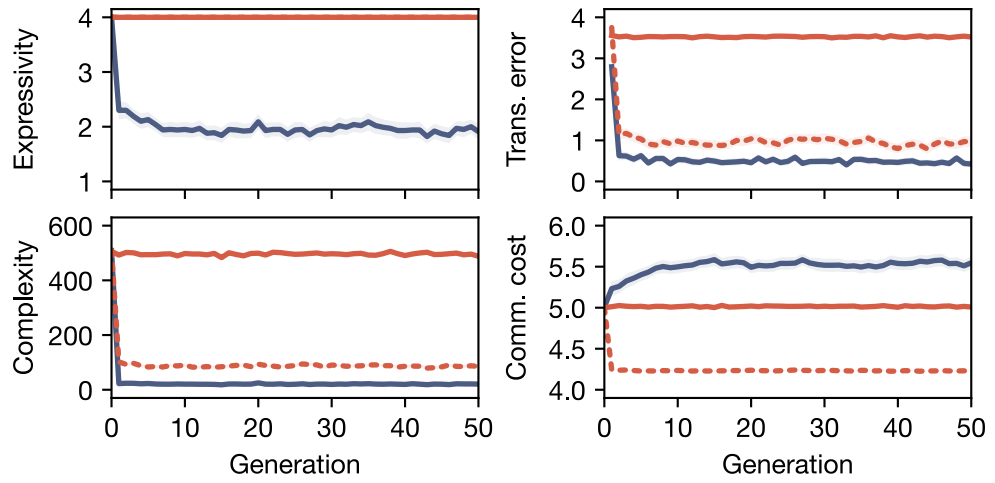
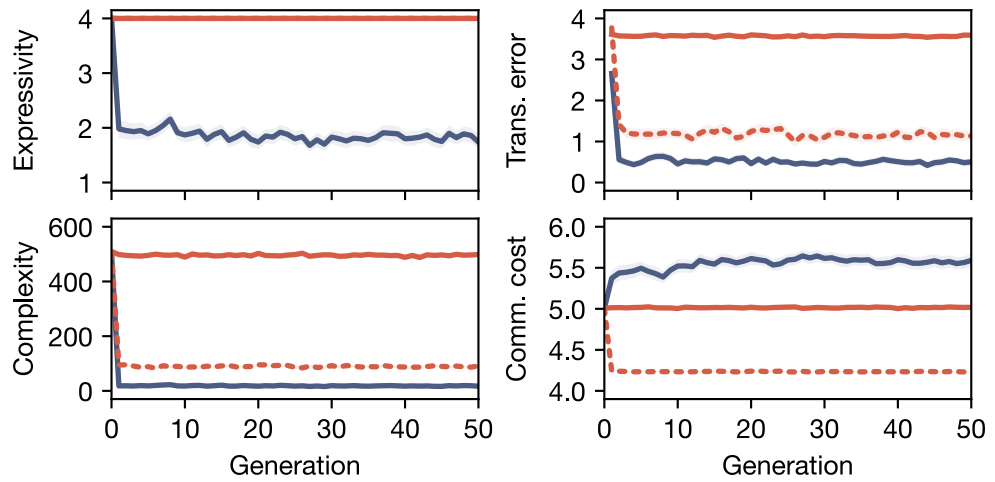
Biology underpins language, but languages themselves are culturally learned, used, and transmitted, and their structural properties adapt in response to various cultural pressures. This thesis has shown that inductive reasoning, under a well-motivated prior bias, acts as a pressure for simple and, ultimately, degenerate languages; this is counteracted, however, by communicative interaction which provides the pressure for informativeness. It is only in the presence of both pressures that structured languages emerge; the more we talk, the more precise our languages become in communicating thought.

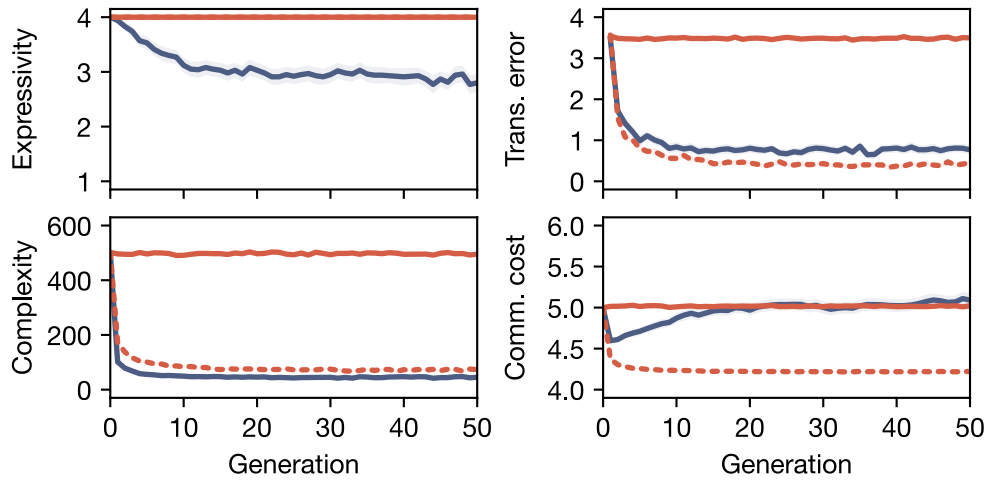
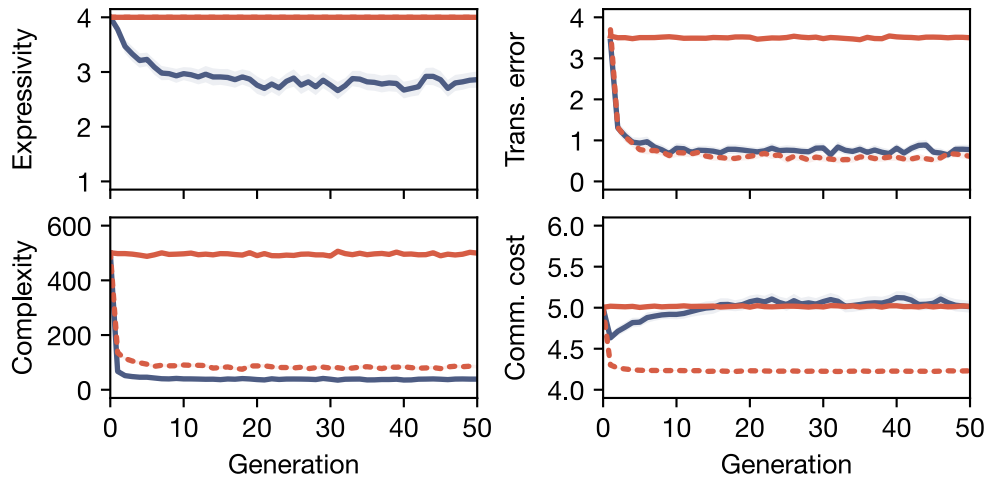
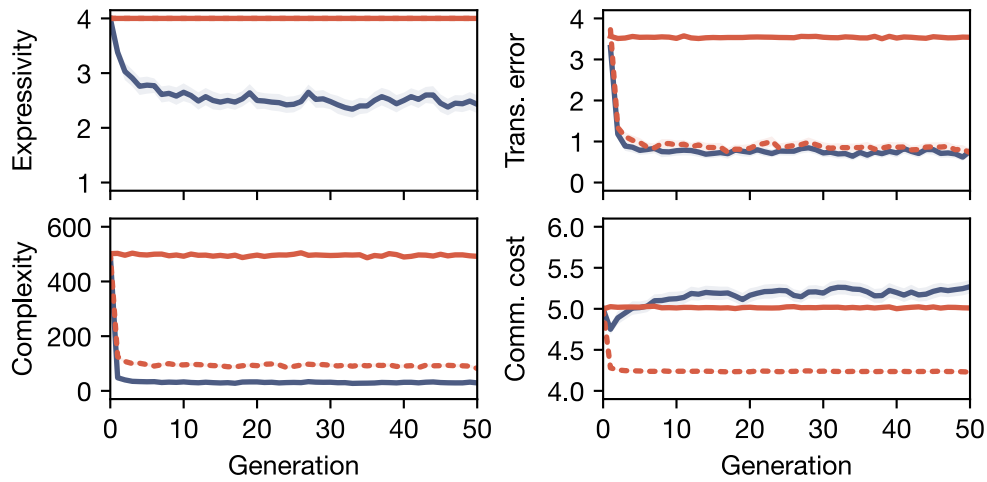


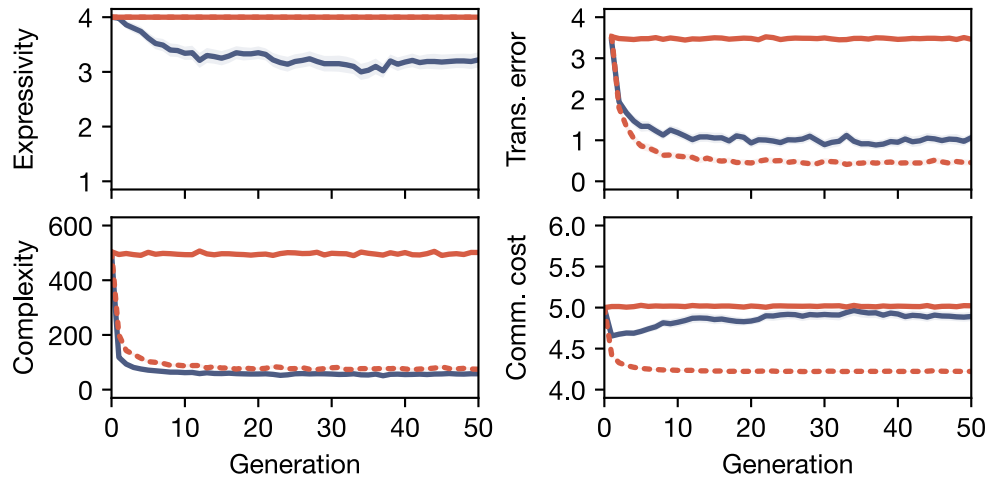
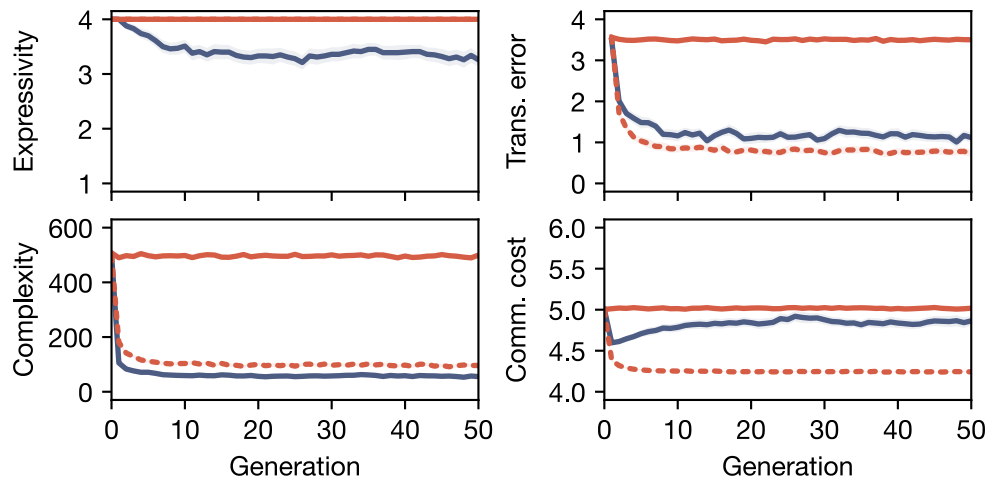
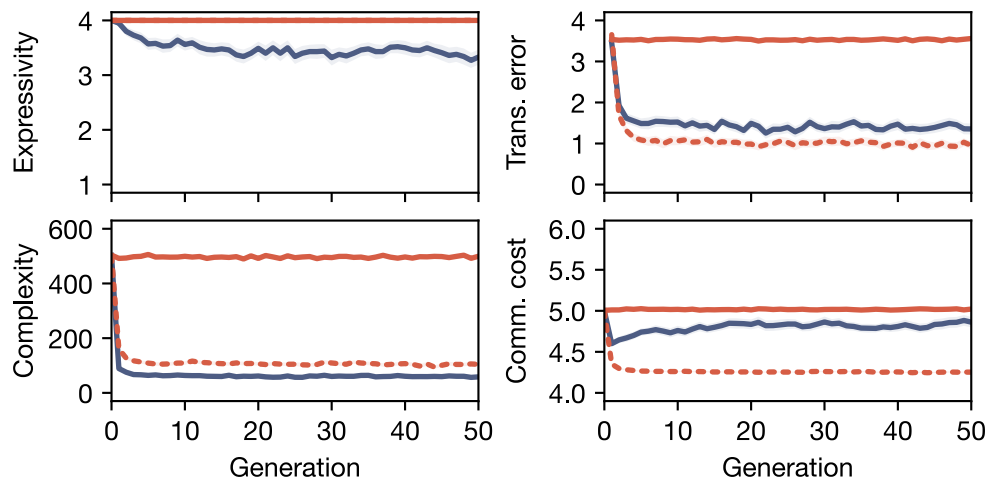
## Appendix A

# Paper 2, Supplement S2: All model results

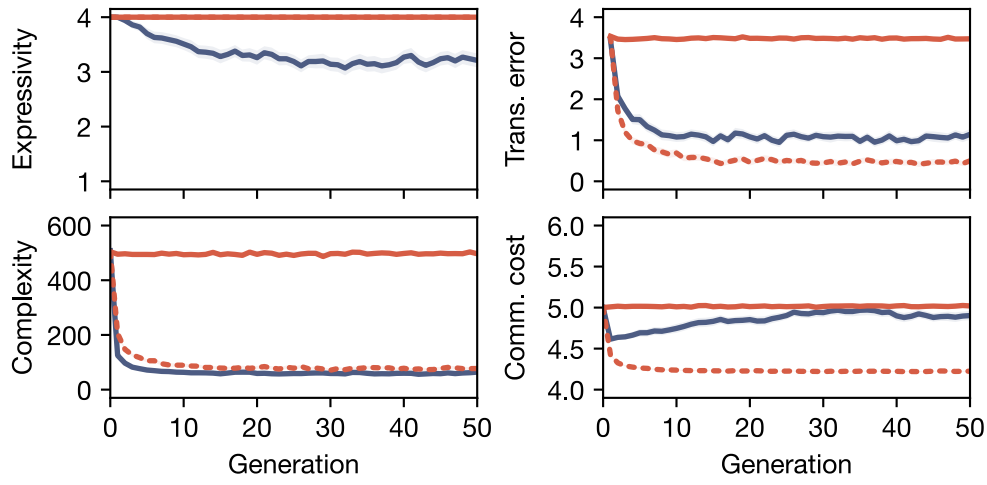
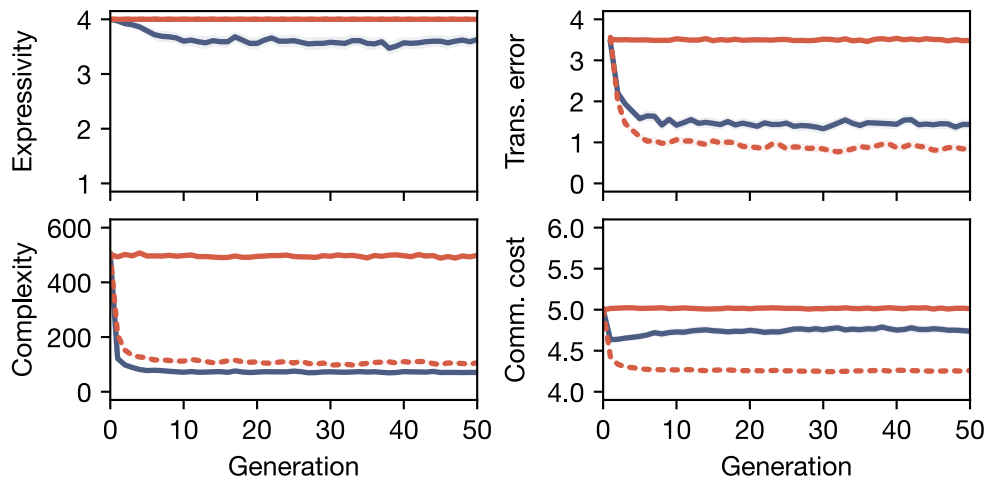
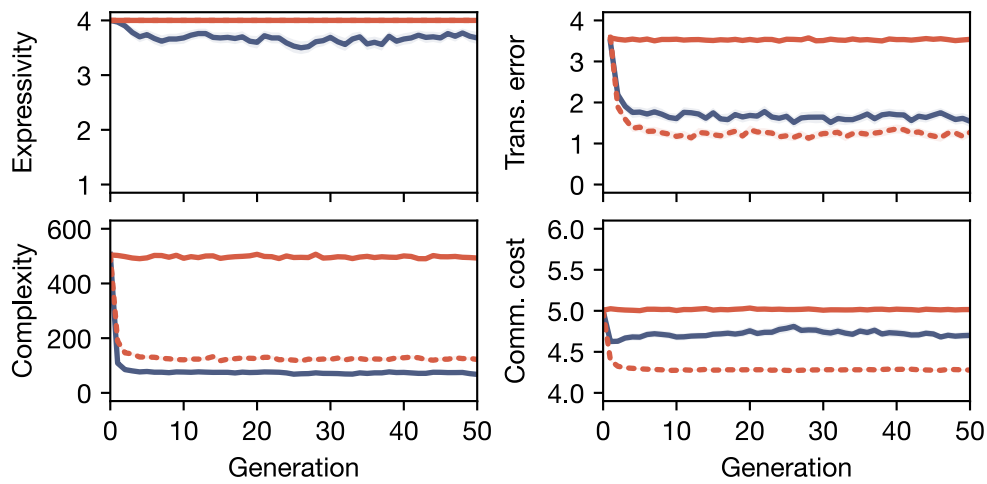
Over the subsequent pages I provide model results under 48 combinations of the bottleneck  $b$ , exposure level  $\xi$ , and noise level  $\varepsilon$ . As in Paper 2, results for expressivity, transmission error, complexity, and communicative cost are given under the three priors (simplicity in blue, informativeness in red, and strong informativeness in dashed red); see pages 64–65 for additional information. Each line represents the mean over 100 chains. The results reveal how greater intergenerational information loss (a tighter bottleneck, fewer exposures, or noisier productions) leads to greater convergence to the prior bias – lower complexity in the case of the simplicity bias and lower cost in the case of the strong informativeness bias. These results can also be explored more interactively at <https://joncarr.net/p/shepard/>

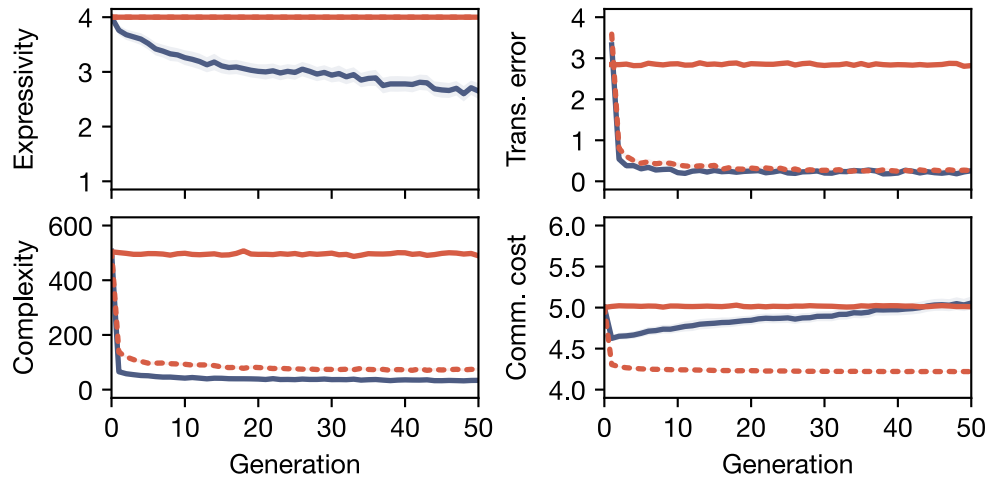
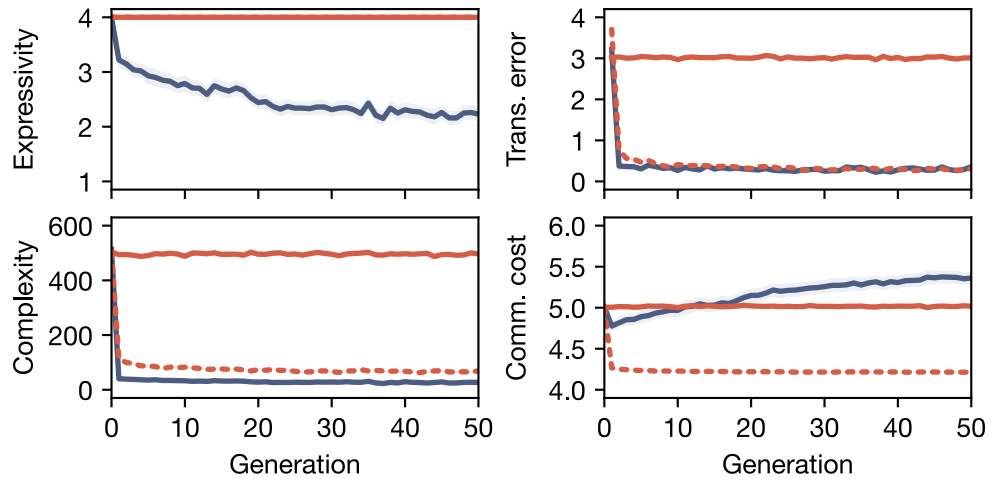
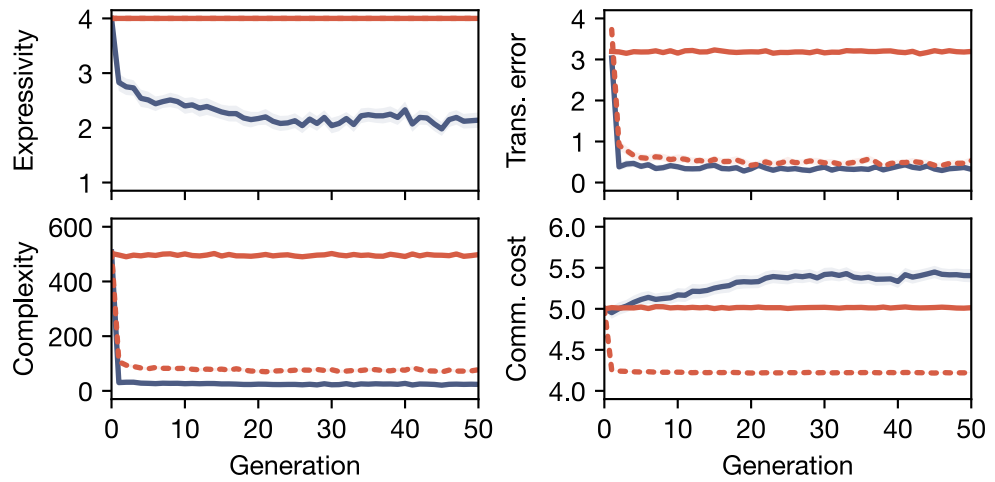
$b = 1, \xi = 1, \varepsilon = 0.01$ 

 $b = 1, \xi = 1, \varepsilon = 0.05$ 

 $b = 1, \xi = 1, \varepsilon = 0.1$ 


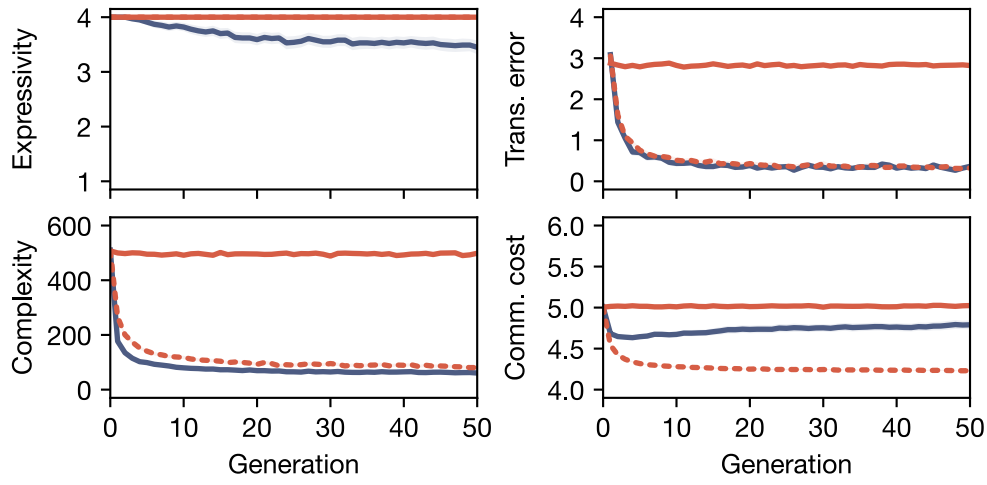
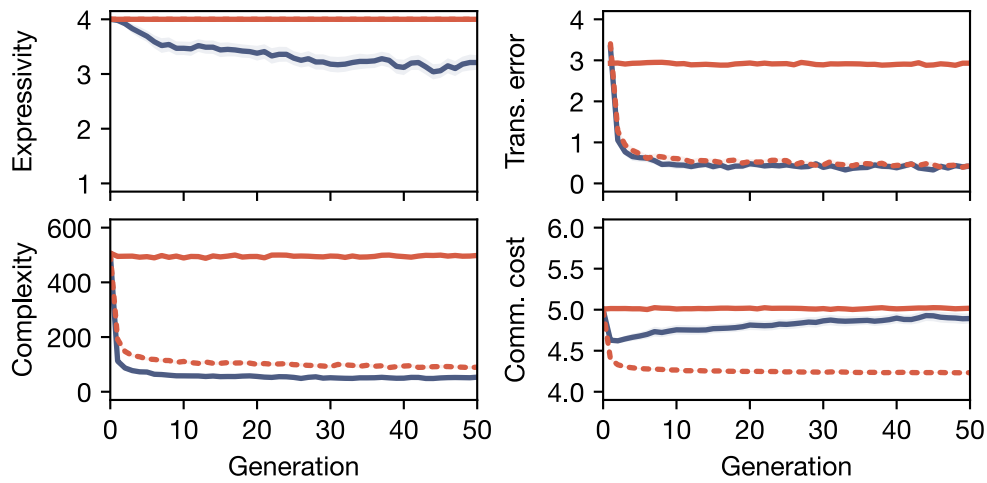
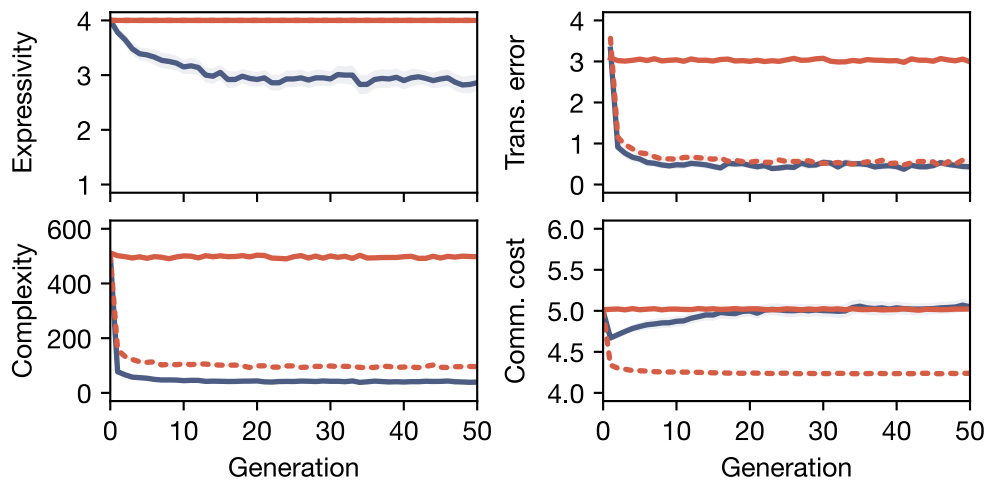
$b = 1, \xi = 2, \varepsilon = 0.01$ 

 $b = 1, \xi = 2, \varepsilon = 0.05$ 

 $b = 1, \xi = 2, \varepsilon = 0.1$ 


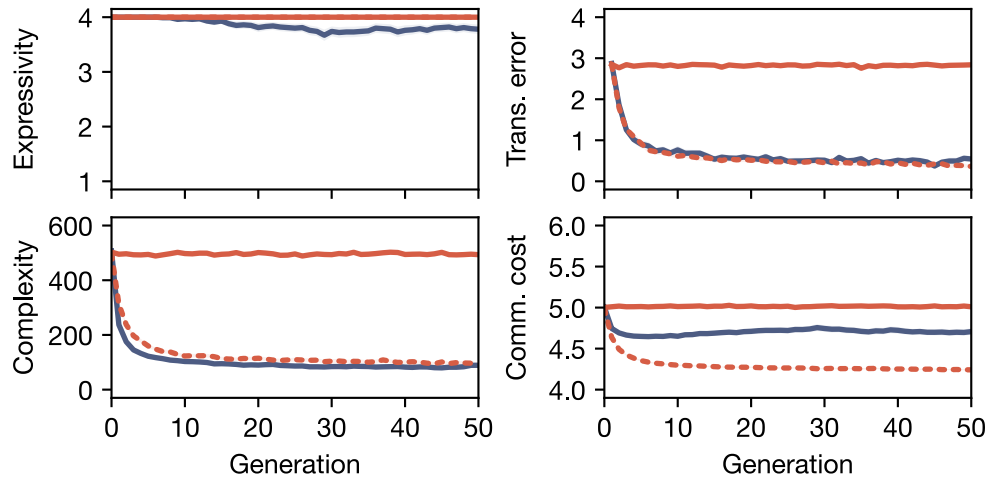
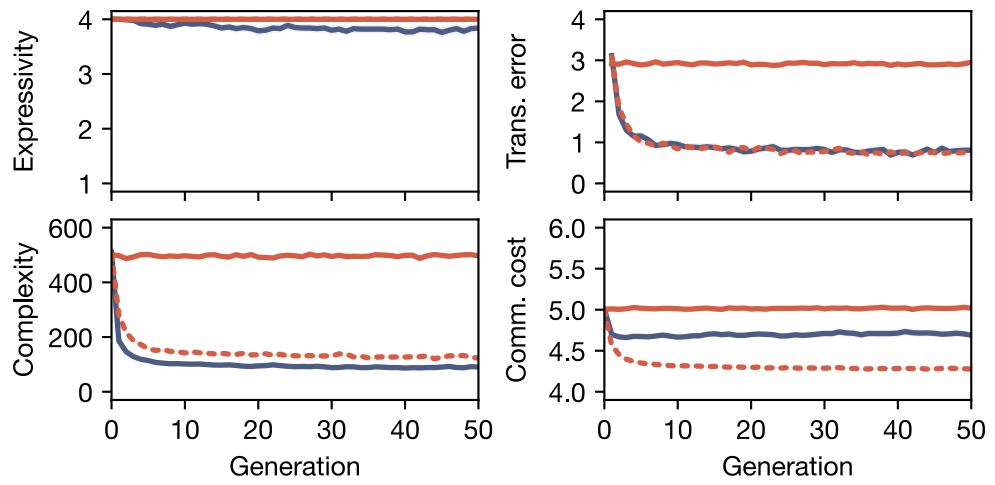
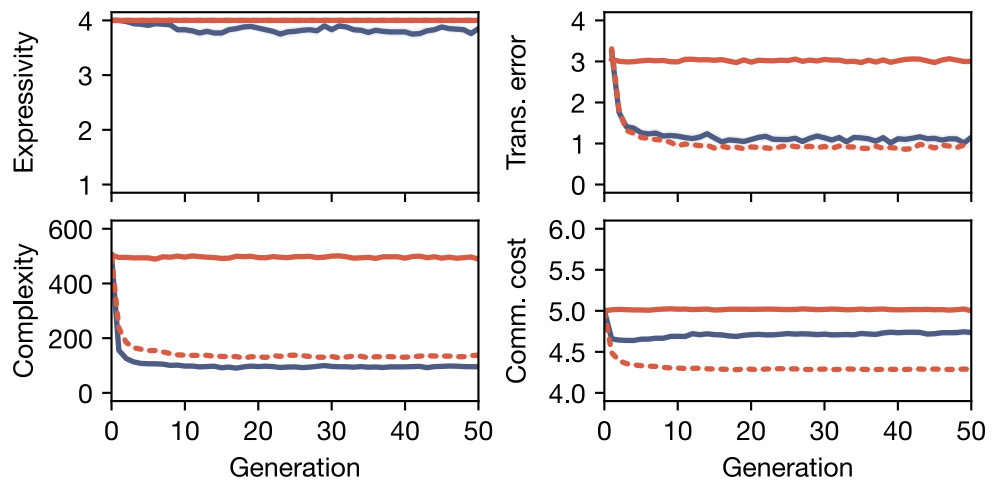
$b = 1, \xi = 3, \varepsilon = 0.01$ 

 $b = 1, \xi = 3, \varepsilon = 0.05$ 

 $b = 1, \xi = 3, \varepsilon = 0.1$ 


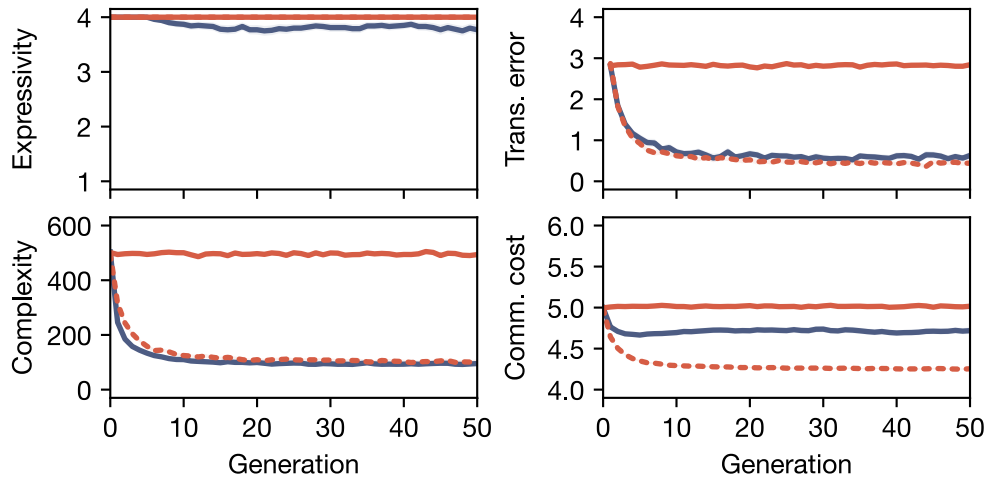
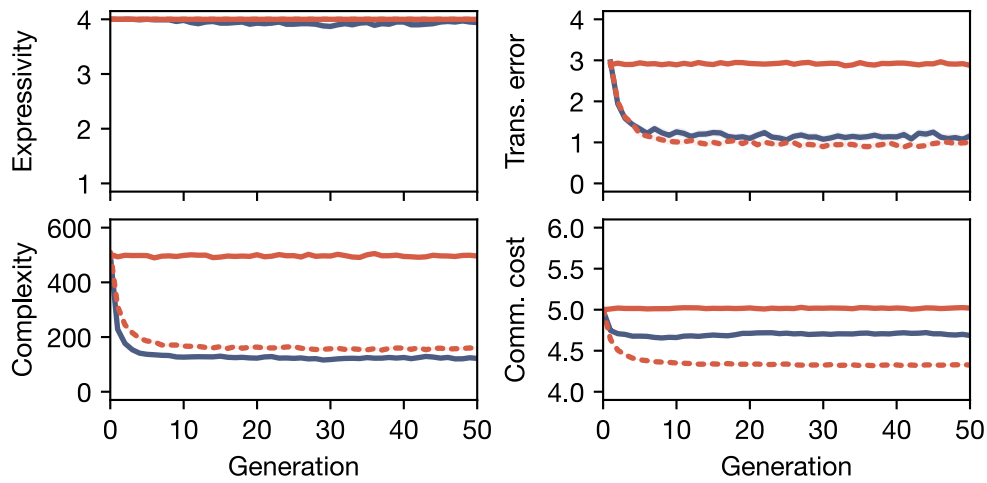
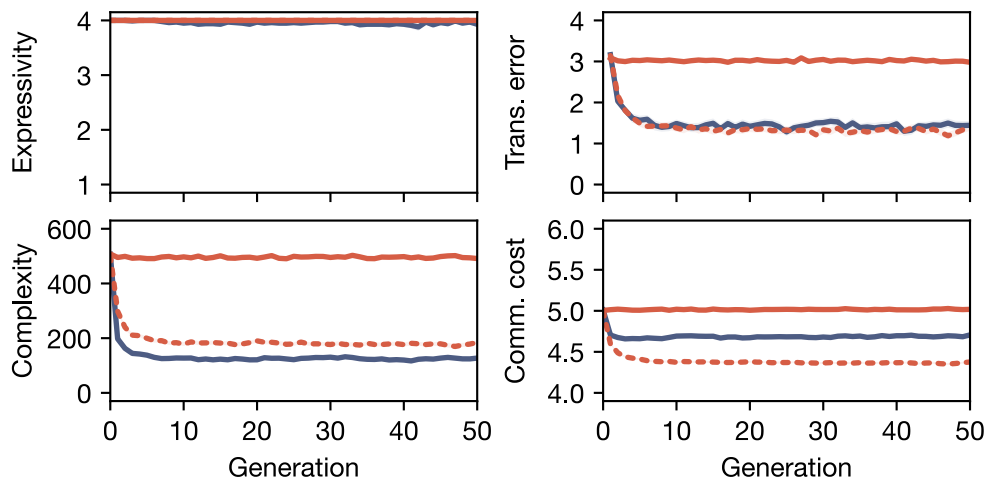


$b = 1, \xi = 4, \varepsilon = 0.01$ 

 $b = 1, \xi = 4, \varepsilon = 0.05$ 

 $b = 1, \xi = 4, \varepsilon = 0.1$ 


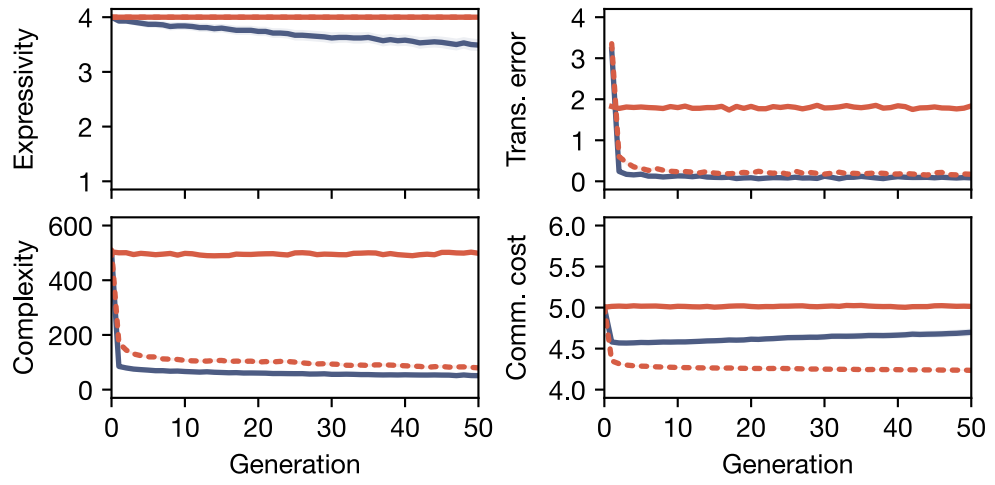
$b = 2, \xi = 1, \varepsilon = 0.01$ 

 $b = 2, \xi = 1, \varepsilon = 0.05$ 

 $b = 2, \xi = 1, \varepsilon = 0.1$ 


$b = 2, \xi = 2, \varepsilon = 0.01$ 

 $b = 2, \xi = 2, \varepsilon = 0.05$ 

 $b = 2, \xi = 2, \varepsilon = 0.1$ 


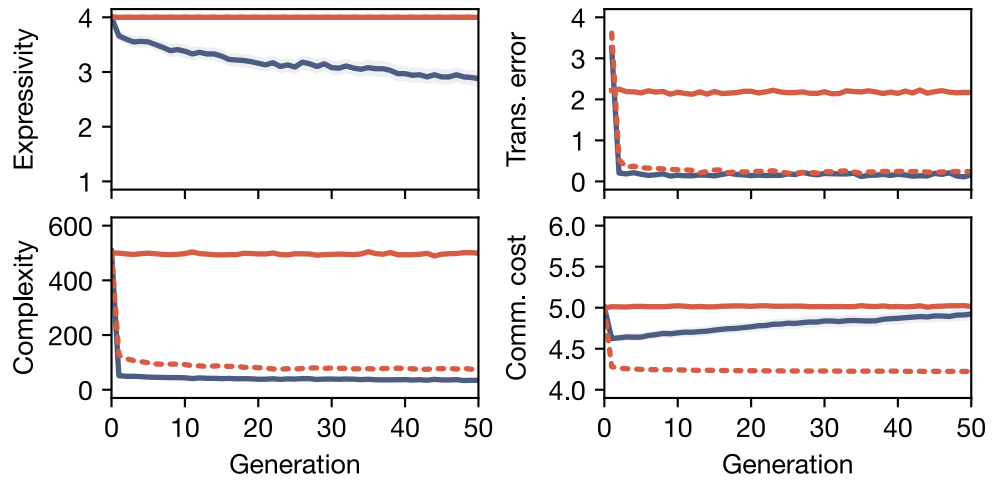
$b = 2, \xi = 3, \varepsilon = 0.01$ 

 $b = 2, \xi = 3, \varepsilon = 0.05$ 

 $b = 2, \xi = 3, \varepsilon = 0.1$ 


$b = 2, \xi = 4, \varepsilon = 0.01$ 

 $b = 2, \xi = 4, \varepsilon = 0.05$ 

 $b = 2, \xi = 4, \varepsilon = 0.1$ 


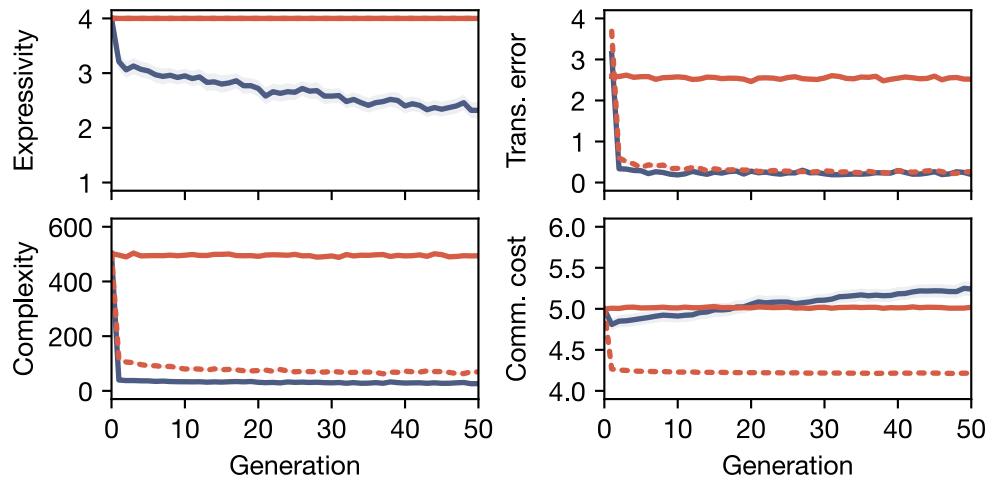
$b = 3, \xi = 1, \varepsilon = 0.01$

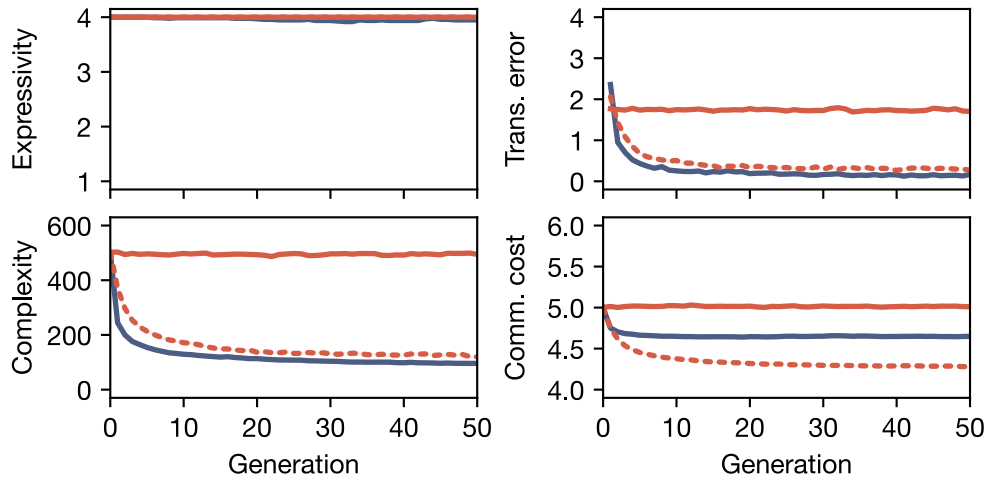
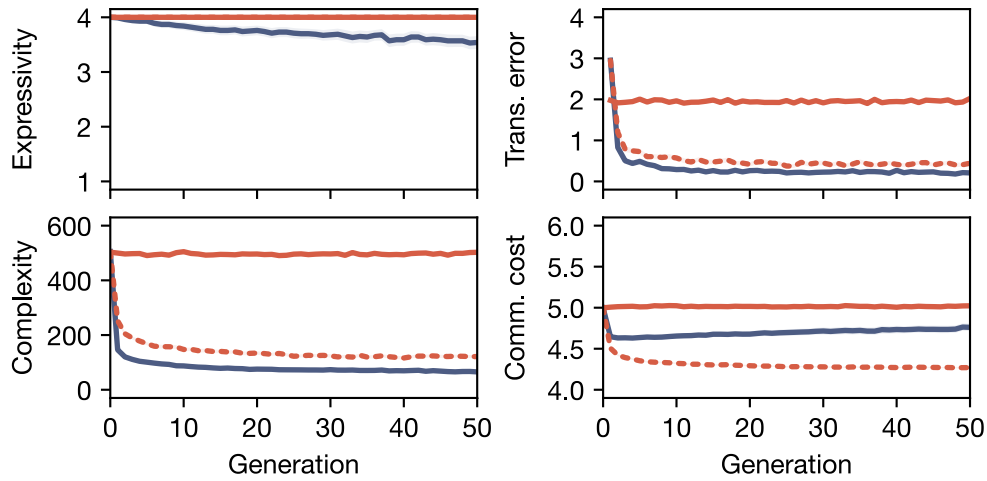
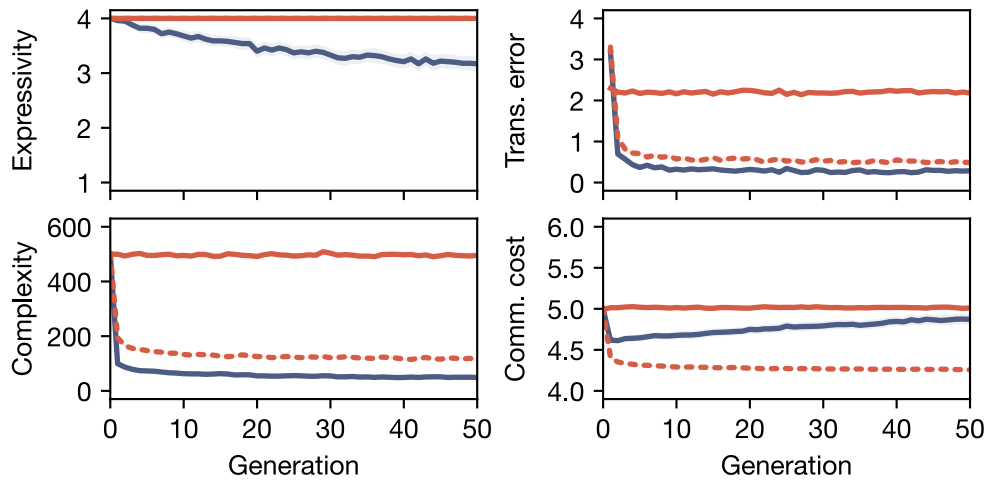


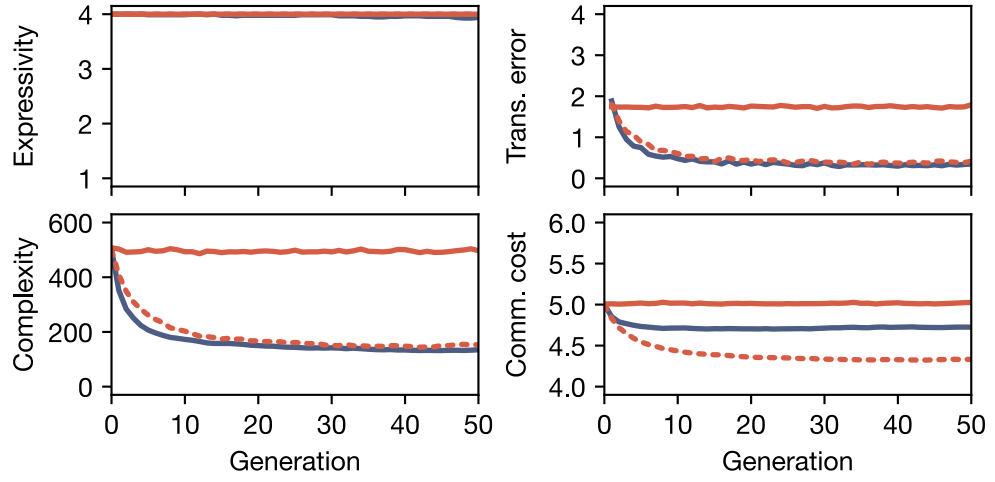
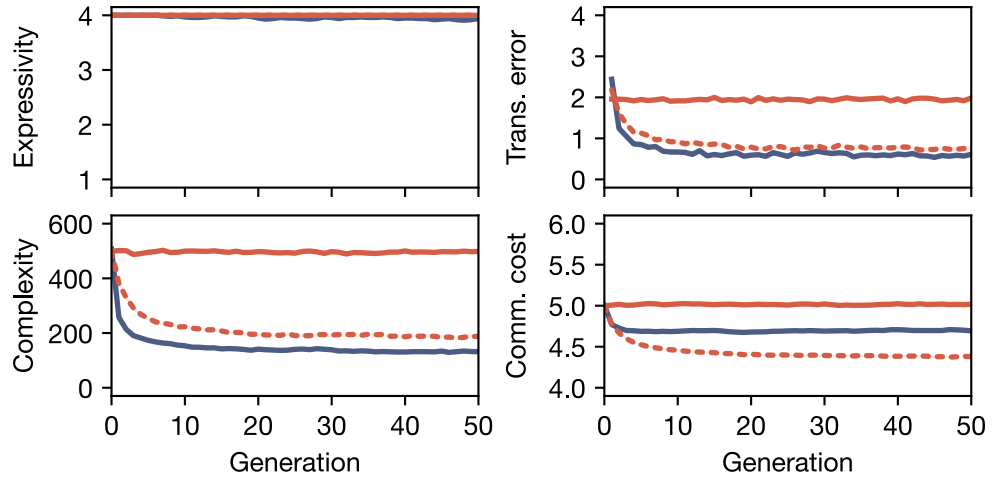
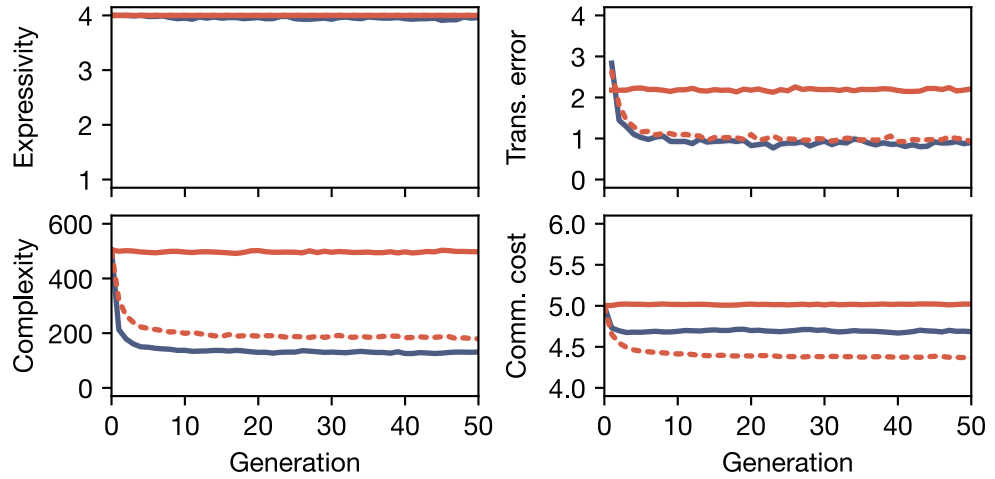
$b = 3, \xi = 1, \varepsilon = 0.05$



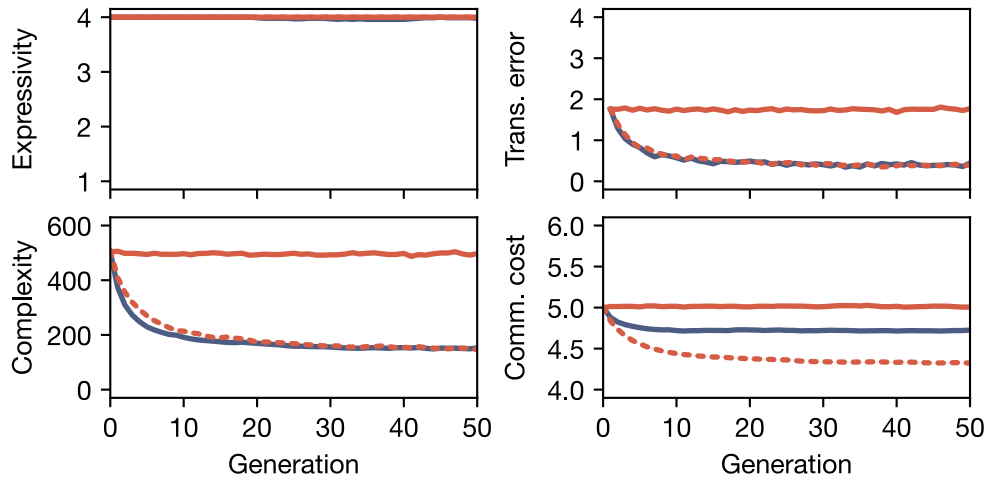
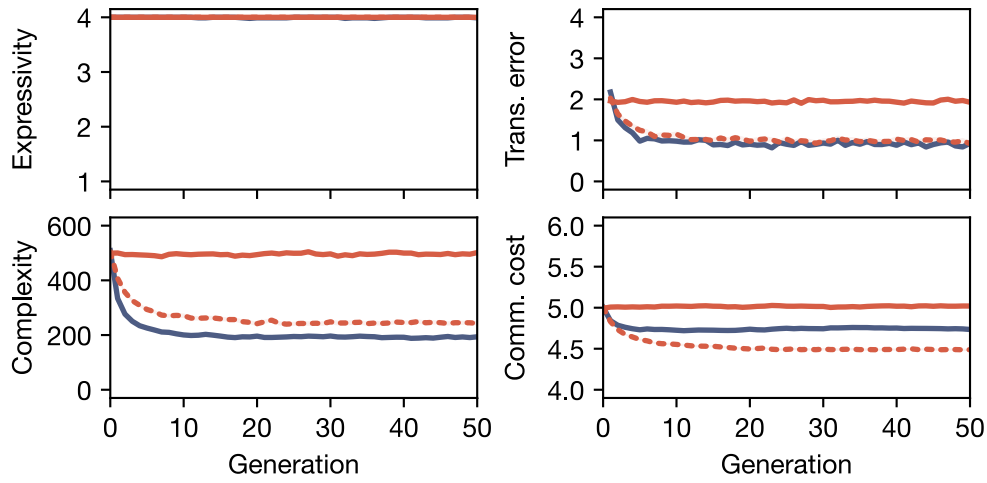
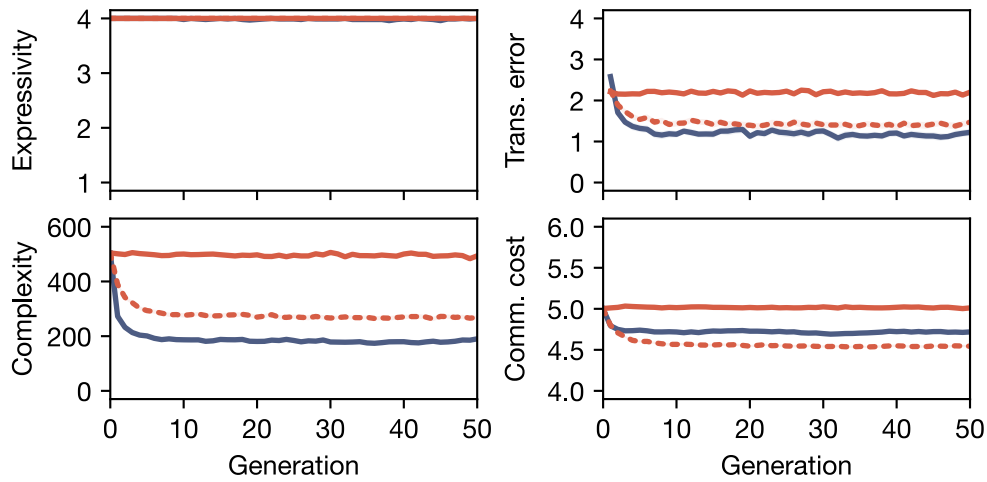
$b = 3, \xi = 1, \varepsilon = 0.1$



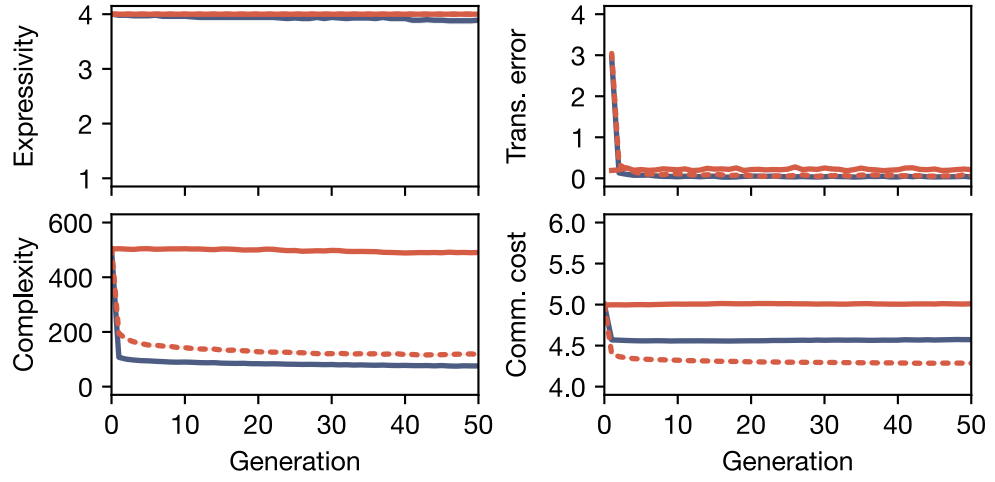
$b = 3, \xi = 2, \varepsilon = 0.01$ 

 $b = 3, \xi = 2, \varepsilon = 0.05$ 

 $b = 3, \xi = 2, \varepsilon = 0.1$ 


$b = 3, \xi = 3, \varepsilon = 0.01$ 

 $b = 3, \xi = 3, \varepsilon = 0.05$ 

 $b = 3, \xi = 3, \varepsilon = 0.1$ 


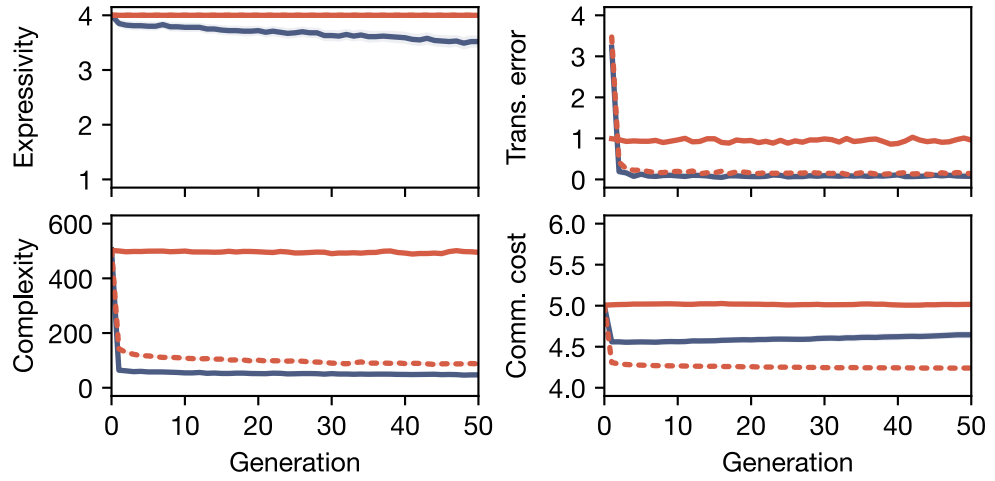


$b = 3, \xi = 4, \varepsilon = 0.01$ 

 $b = 3, \xi = 4, \varepsilon = 0.05$ 

 $b = 3, \xi = 4, \varepsilon = 0.1$ 


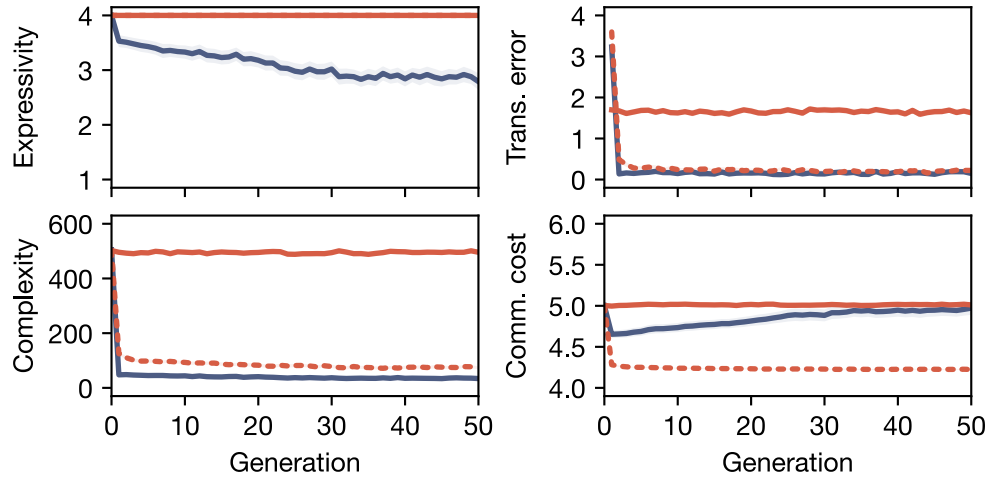
$b = 4, \xi = 1, \varepsilon = 0.01$

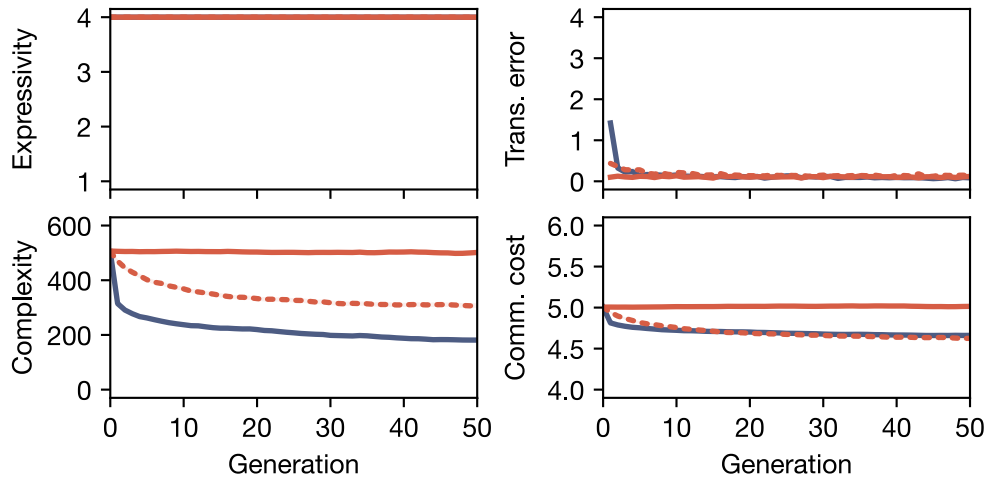
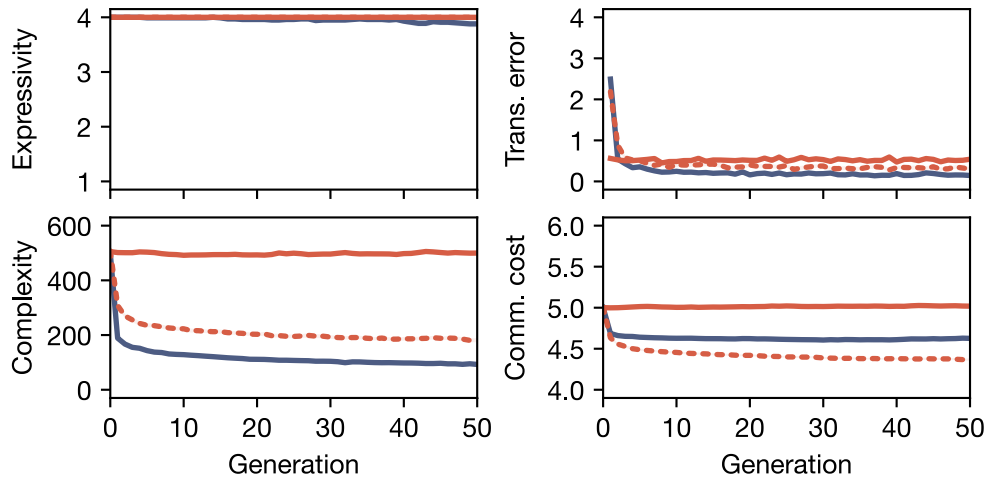
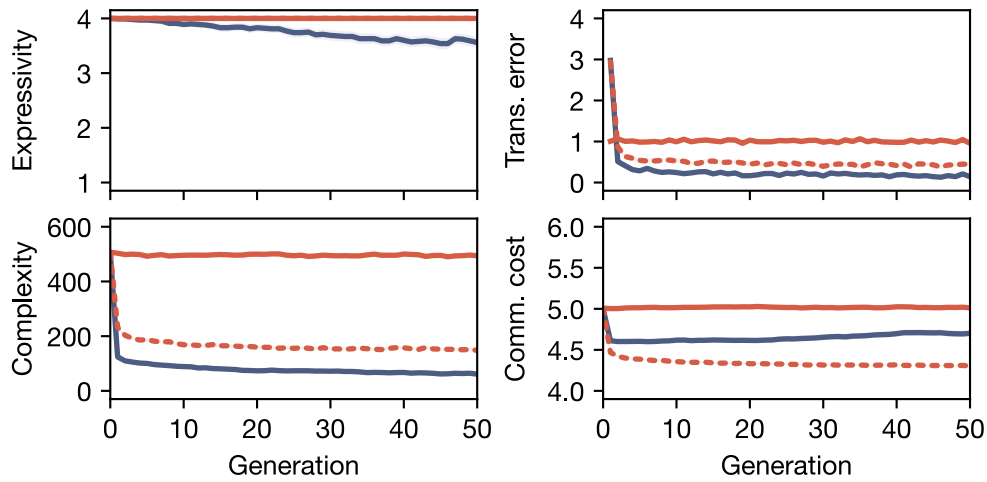


$b = 4, \xi = 1, \varepsilon = 0.05$

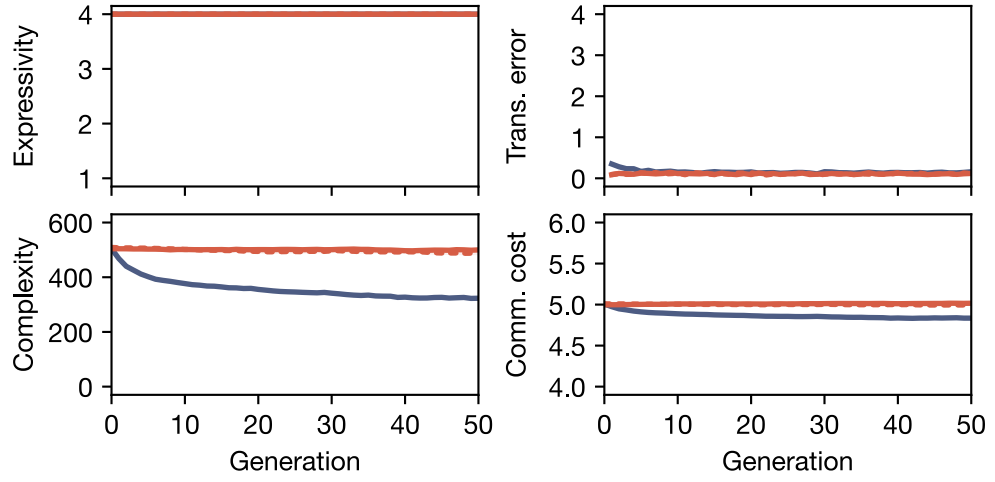


$b = 4, \xi = 1, \varepsilon = 0.1$

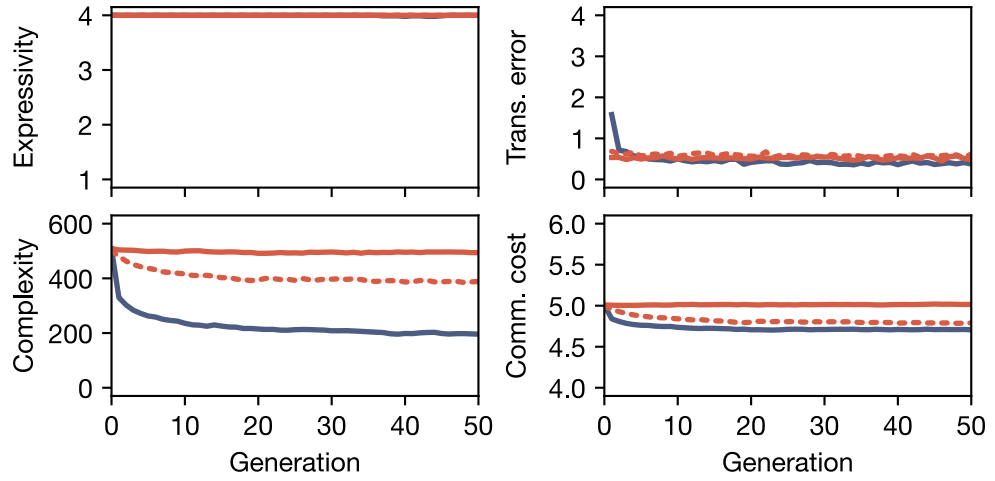


$b = 4, \xi = 2, \varepsilon = 0.01$ 

 $b = 4, \xi = 2, \varepsilon = 0.05$ 

 $b = 4, \xi = 2, \varepsilon = 0.1$ 


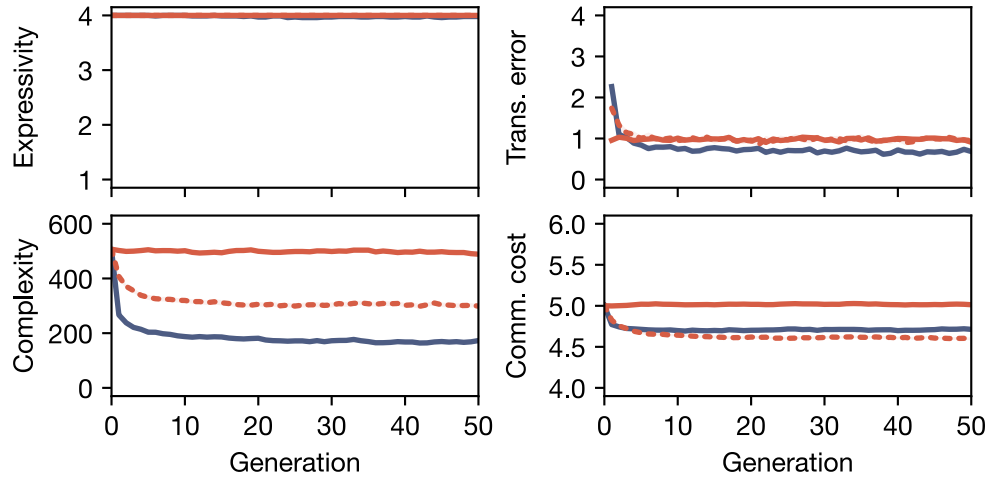
$b = 4, \xi = 3, \varepsilon = 0.01$

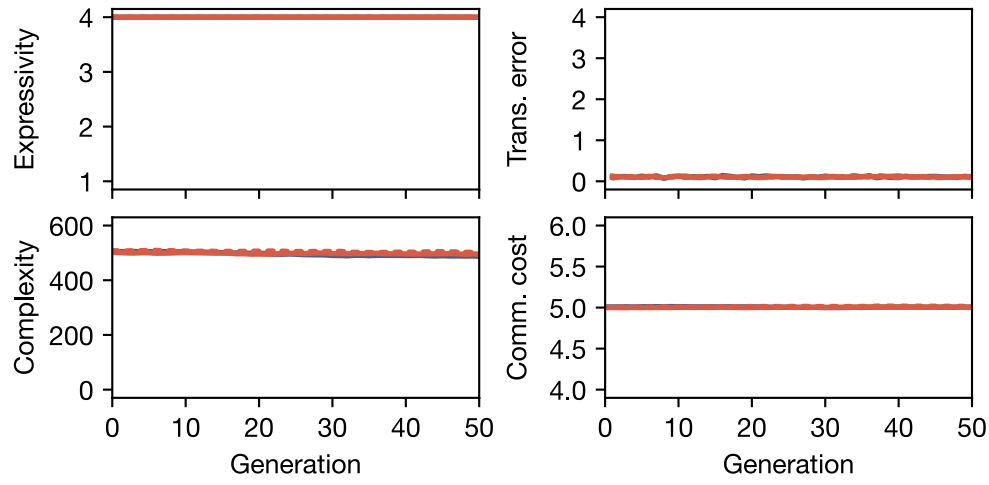
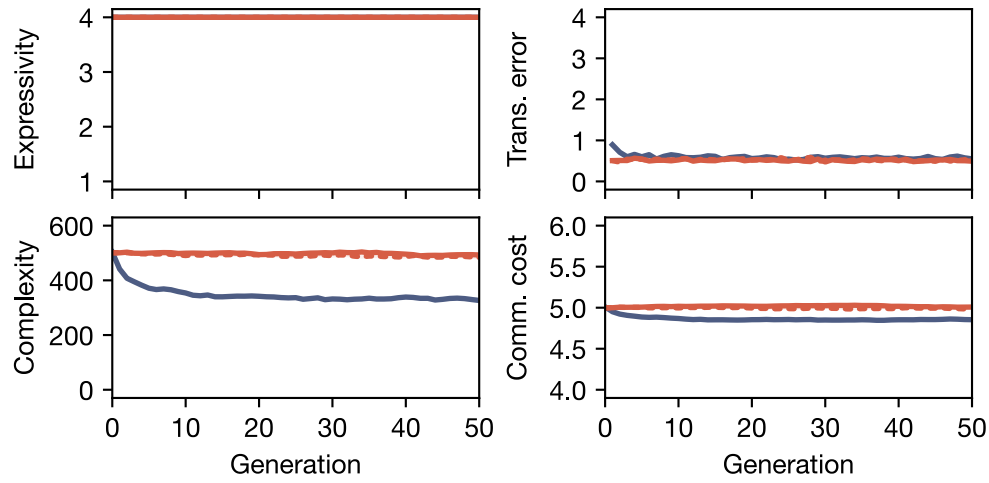
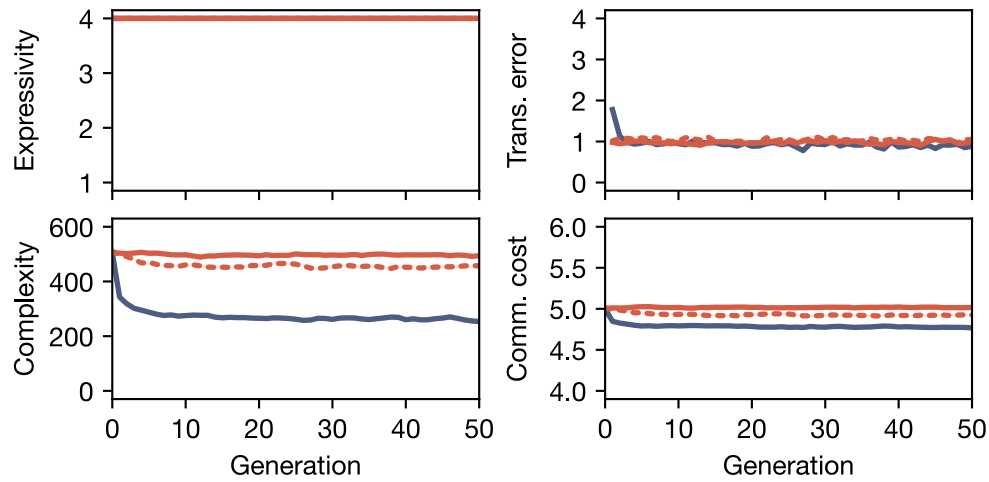


$b = 4, \xi = 3, \varepsilon = 0.05$



$b = 4, \xi = 3, \varepsilon = 0.1$



$b = 4, \xi = 4, \varepsilon = 0.01$ 

 $b = 4, \xi = 4, \varepsilon = 0.05$ 

 $b = 4, \xi = 4, \varepsilon = 0.1$ 


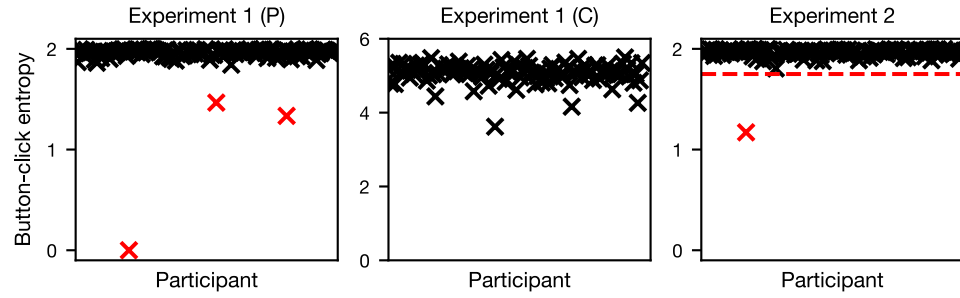


## Appendix B

### Paper 2, Supplement S3: Participant exclusion and attrition

On every trial, the participant had to click one of four response buttons (or one of 64 response stimuli in the case of the Comprehension condition in Experiment 1). The mapping between labels/stimuli and buttons was randomized on every trial, such that the participant would have to look over the buttons to find the one they wanted. Therefore, we would expect to find that *buttons* (but not necessarily labels/stimuli) are being clicked at random. If button clicks appear to be nonrandom, this would suggest that the participant is repeatedly clicking the same button without regard to the label (or stimulus). To check for this, I looked at the entropy of button clicks (see Fig. B.1). Random button clicking would result in entropy of  $\log 4 = 2$  bits (Production) or  $\log 64 = 6$  bits (Comprehension).

Under the Production test type of Experiment 1, there were three clear outliers (highlighted in red), so these participants were excluded and new participants were recruited to take their place (all three had been assigned to the Size-only category system). Under the Comprehension test type, we might expect to find a little less randomness in button clicks; participants might hover their cursor around one particular area of the stimulus picker and click the closest stimulus that belongs to the target category, resulting in some buttons being used more frequently than others. Since the results indicated no strong outliers, we retained all participants. In Experiment 2, which is production-based, we preset the exclusion criterion to 1.75 bits (shown by the dashed



**Figure B.1:** Button-click entropy of participants in Experiment 1 (Production and Comprehension) and Experiment 2. Each cross is an individual participant. A total of four participants were excluded, highlighted in red.

**Table B.1:** Participant numbers by condition in Experiment 1 (recruited – terminated – excluded)

	Production	Comprehension	TOTAL
<b>Angle-only</b>	44 – 4 – 0 = 40	46 – 6 – 0 = 40	90 – 10 – 0 = 80
<b>Size-only</b>	50 – 7 – 3 = 40	70 – 30 – 0 = 40	120 – 37 – 3 = 80
<b>Angle &amp; Size</b>	41 – 1 – 0 = 40	58 – 18 – 0 = 40	99 – 19 – 0 = 80
<b>TOTAL</b>	135 – 12 – 3 = 120	174 – 54 – 0 = 120	309 – 66 – 3 = 240

red line in Fig. B.1), which was set based on the Production results of Experiment 1. Any participant whose button-click entropy was below this value was excluded automatically and a new participant was automatically recruited to fill that generation in the chain. This was applied in just one instance (highlighted in red). Therefore, a total of four participants (0.85%) were excluded across the two Experiments. Three from the Production/Size-only condition and one from the iterated learning experiment.

It has been shown that online experiments may be adversely affected by high participant attrition (i.e. termination of the experiment before it is completed), especially where attrition may be linked to experimental condition (Zhou & Fishbach, 2016). In our experiments, for example, participants may be more likely to terminate the experiment if they are assigned to a condition where the system is harder to learn, which could have the effect of making difficult-to-learn systems appear easier than they actually are.

A total of 309 participants began Experiment 1. Of these, 66 terminated the experiment prior to completing it, so their data were erased because they were deemed to have withdrawn consent. Participant numbers by condition are presented in Table B.1. Participants who terminated the experiment were disproportionately likely to have been assigned to the Size-only system or the Comprehension test type. This means that our



results may, for example, overestimate how easy the Size-only system was to learn, under the assumption that participants are more likely to terminate a task if they find it difficult. A total of 273 participants began Experiment 2. Of these, 48 participants terminated the experiment prior to completing it. Participants who decided to terminate either experiment had the facility to leave a comment explaining why, but none availed of this, making it difficult to establish their motivations. On average, participants terminated the experiments after around 3 minutes, and around 40% of them did not progress beyond the instructions page.

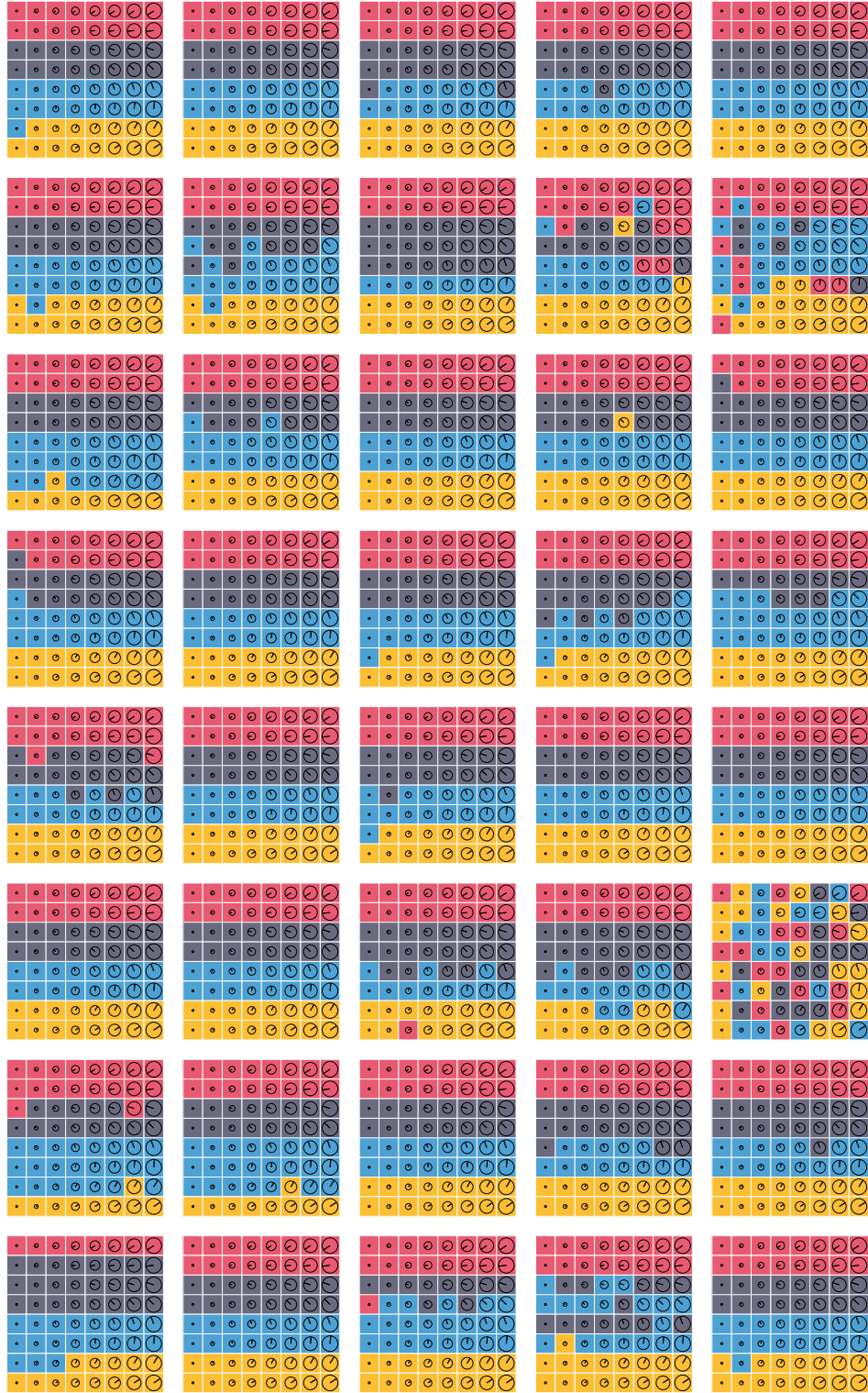


## Appendix C

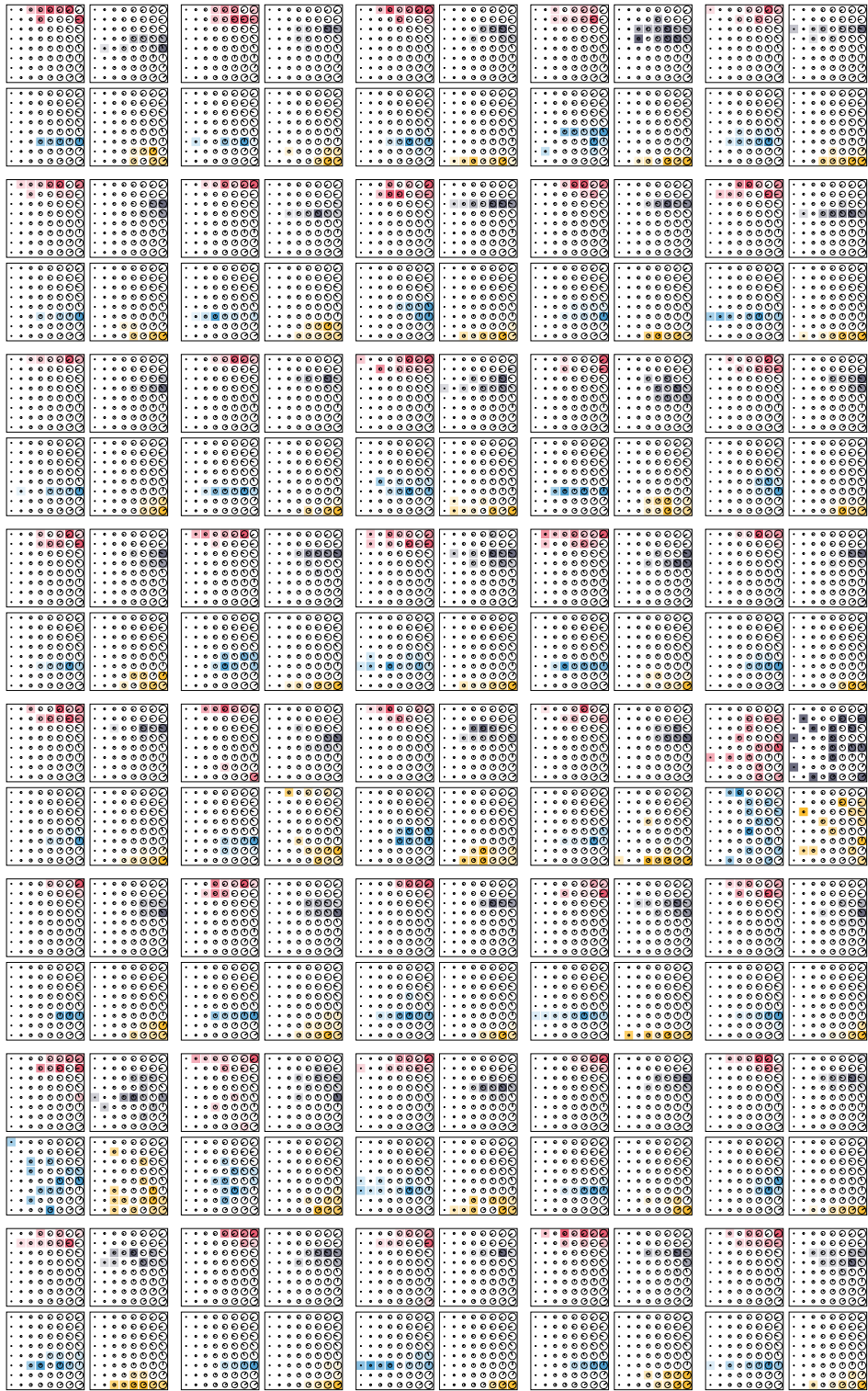
# Paper 2, Supplement S4: Individual participant results in Experiment 1

Over the subsequent pages, I provide categorization results from each of the 240 participants who took part in Experiment 1 (see pages 67–72). Each page gives the results from the 40 participants in a given condition. Participants were assigned to one of three category systems (Angle-only, Size-only, or Angle & Size) and one of two test types (Production or Comprehension). Colours indicate the category a particular stimulus was assigned to by the participant. In the case of the Comprehension results, each participant's results are separated out across four grids and the lightness of the colour indicates how many times that stimulus was selected as an example of the category. In general, participants learning the Angle-only system reproduced that system very accurately; participants learning the Size-only system were less accurate; and participants learning the Angle & Size system had quite low accuracy.

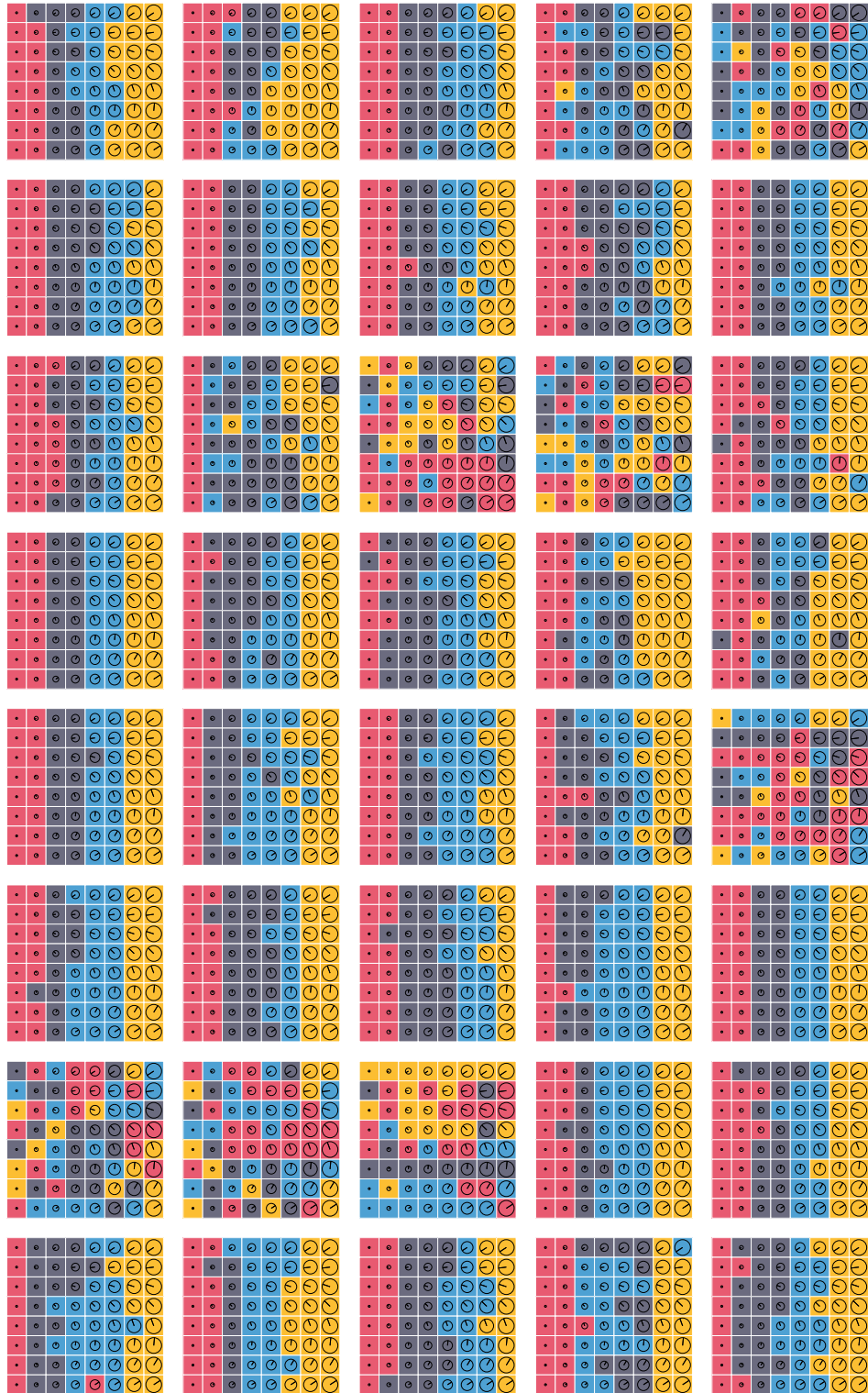
## Angle-only / Production



Angle-only / Comprehension

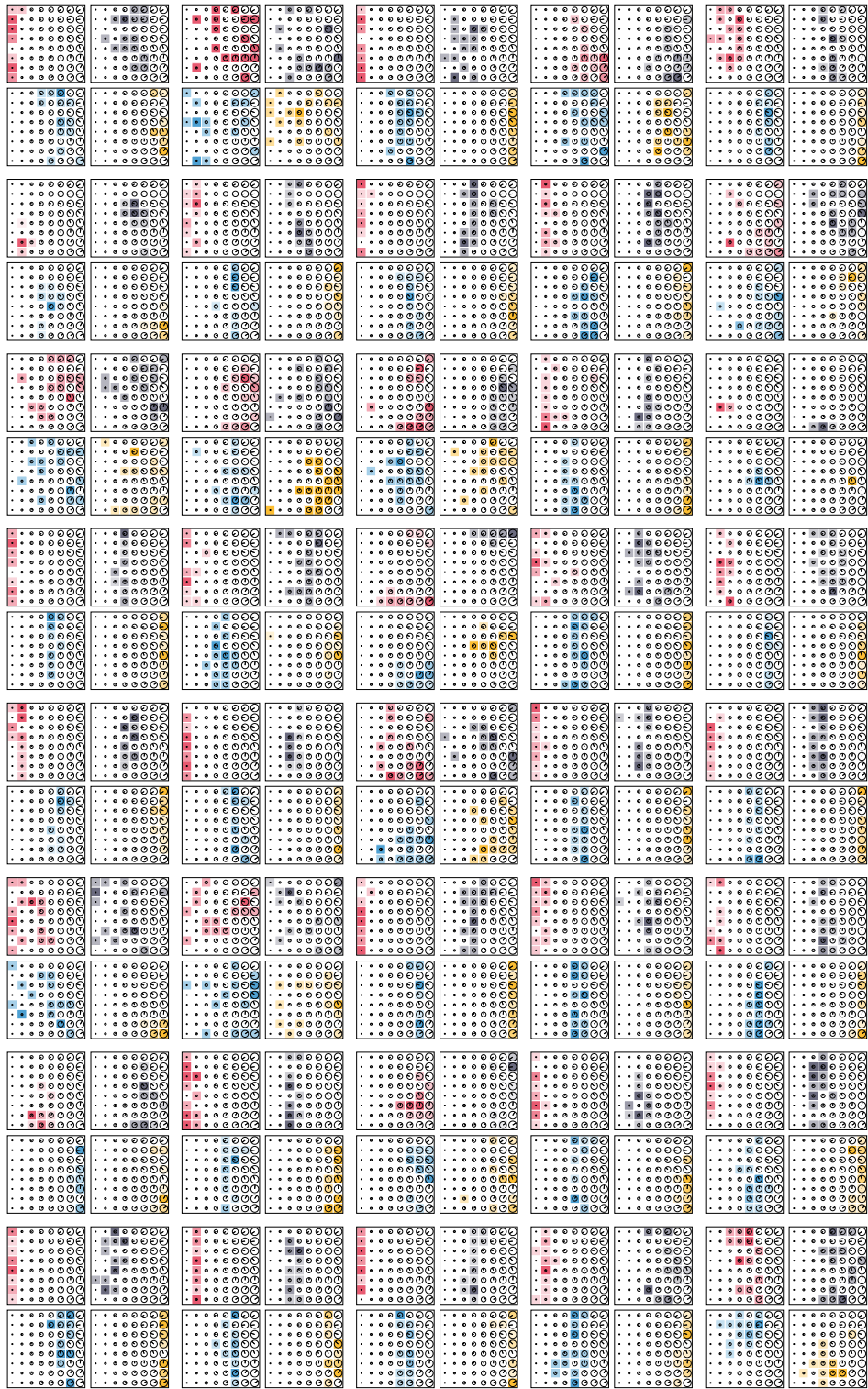


## Size-only / Production

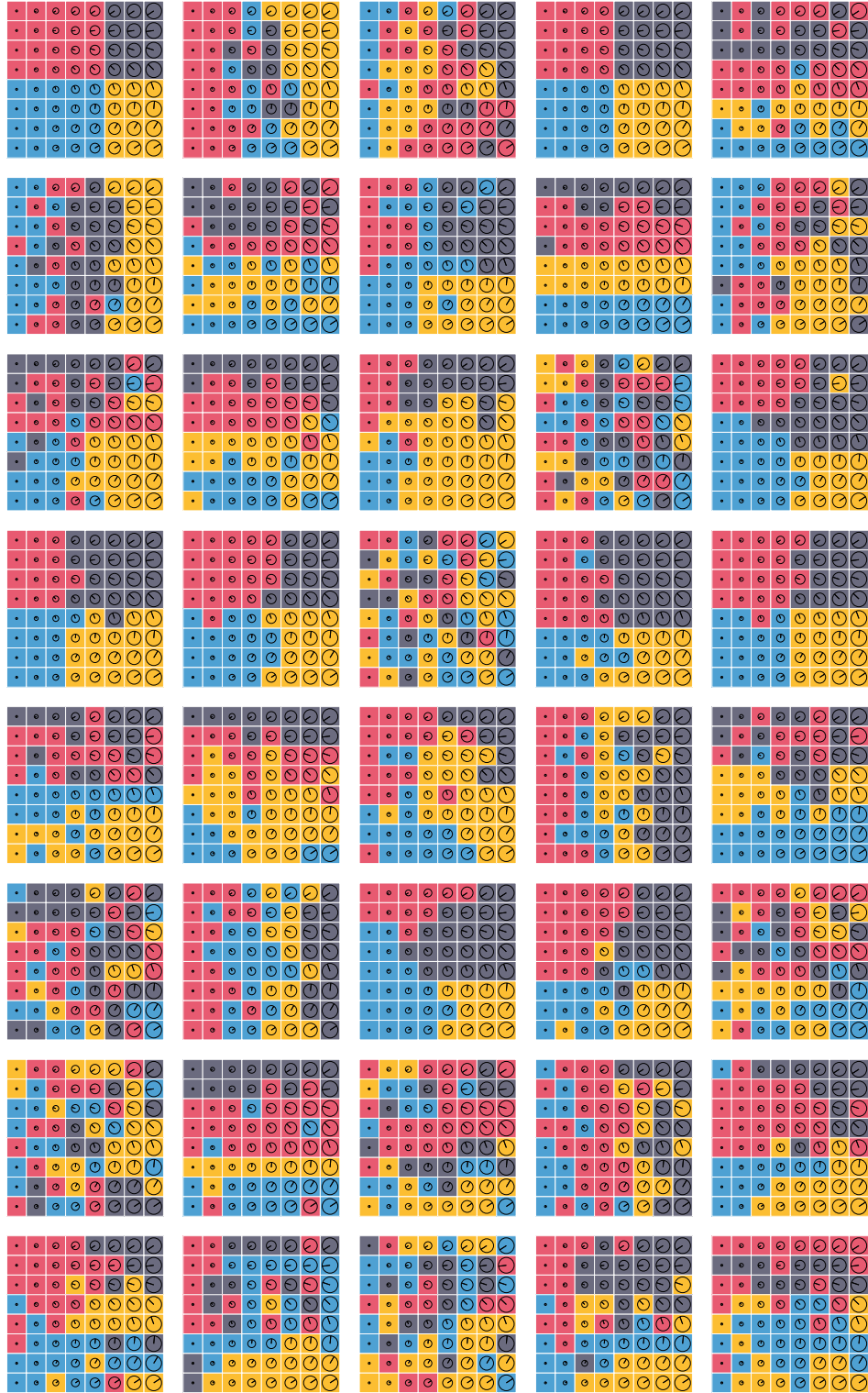




Size-only / Comprehension

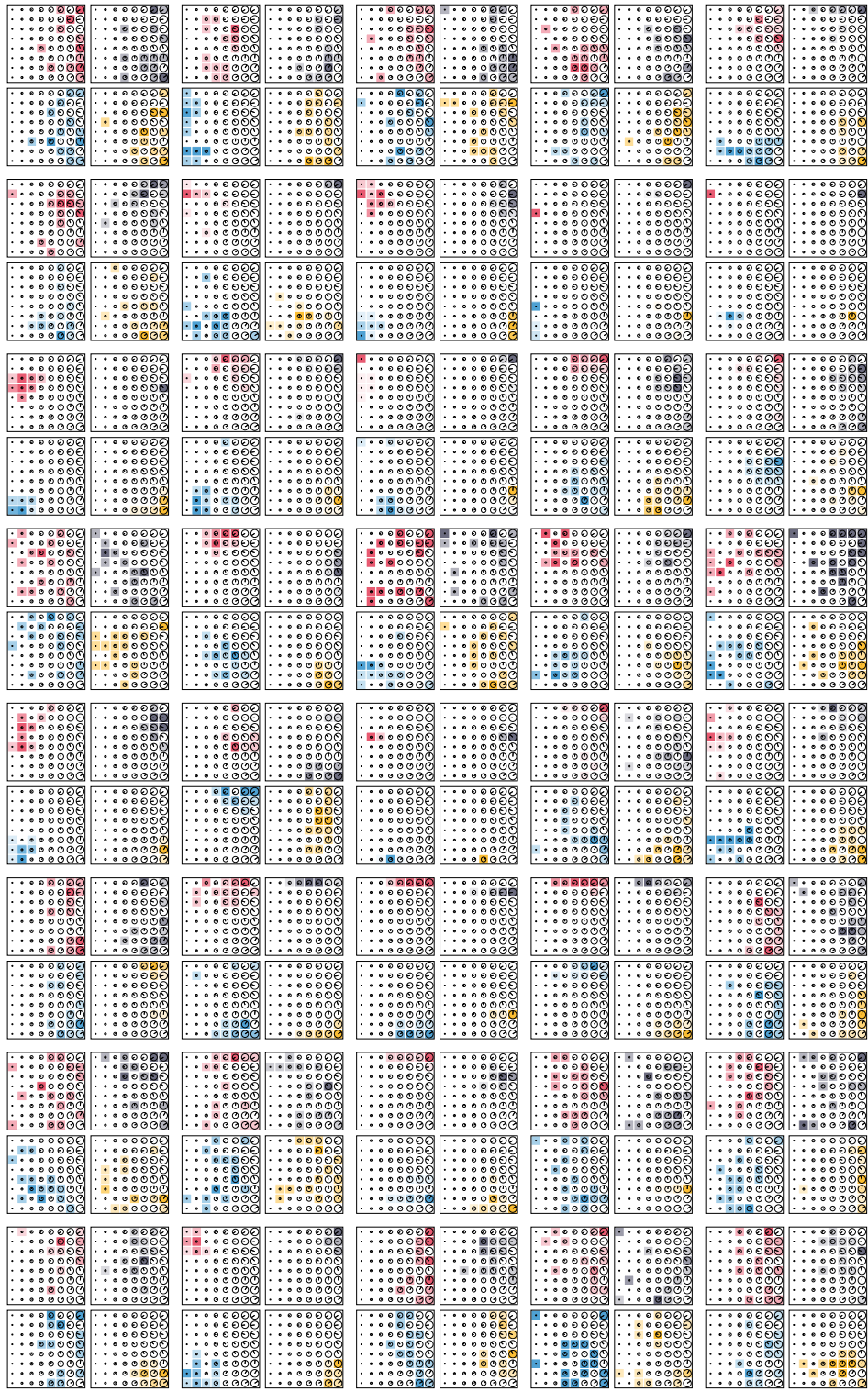


## Angle &amp; Size / Production





Angle & Size / Comprehension





## **Appendix D**

### **Paper 3, Supplement S1:**

### **Experimental briefs**

Over the following three pages, I provide the briefing materials that were given to participants in each of the three experiments reported in Paper 3. This briefing information was accompanied by oral explanation of the task, and participants has the opportunity to ask questions.

# Flatlanders experiment



## Brief

You have just entered a parallel universe that has only two dimensions! This curious place is inhabited by an intelligent life form, the Flatlanders, who are obsessed with two-dimensional shapes and have a huge vocabulary just for triangles alone.

Your task is to learn the words that the Flatlanders use for triangles to help us establish contact with these strange beings. It's a pretty difficult task — but we think you're the right person for the job!

### **Stage 1: Training**

You will see a series of triangles, one by one. Each triangle will be presented with its name in the Flatlander language. The name will also be pronounced by the computer to help you learn it. After every third triangle, you will see one of those three triangles again and you must type in its name. This stage is designed to help you learn the language.

### **Stage 2: Test**

Again, you will see a series of triangles. For each triangle, simply type in what you think it's called based on the training you completed in stage 1. The test is designed to assess how well you've learned the Flatlander language, and there's a £20 Amazon voucher for whoever learns it best.

You will learn a lot of words very quickly, and it may be difficult to take it all in. But don't panic! The most important thing is to maintain good relations with the Flatlanders by giving it your best shot. You must type in an answer for every triangle, but it's okay to guess if you're unsure. Even if you get the word wrong, you'll still get points for getting the word partially correct.

Good luck!

# Flatlanders experiment



## Brief

You have just entered a parallel universe that has only two dimensions! This curious place is inhabited by an intelligent life form, the Flatlanders, who are obsessed with two-dimensional shapes and have a huge vocabulary just for triangles alone.

Your task is to learn the words that the Flatlanders use for triangles to help us establish contact with these strange beings. It's a pretty difficult task — but we think you're the right person for the job!

### Stage 1: Training

You will see a series of triangles, one by one. Each triangle will be presented with its name in the Flatlander language. The name will also be pronounced by the computer to help you learn it. After every third triangle, you will see one of those three triangles again and you must type in its name. This stage is designed to help you learn the language.

### Stage 2: Test

Again, you will see a series of triangles. For each triangle, simply type in what you think it's called based on the training you completed in stage 1. However, if you use the same word too many times, you will see a message asking you to use a different word. The test is designed to assess how well you've learned the Flatlander language, and there's a £20 Amazon voucher for whoever learns it best.

You will learn a lot of words very quickly, and it may be difficult to take it all in. But don't panic! The most important thing is to maintain good relations with the Flatlanders by giving it your best shot. You must type in an answer for every triangle, but it's okay to guess if you're unsure. Even if you get the word wrong, you'll still get points for getting the word partially correct.

Good luck!

# Flatlanders experiment



## Brief

You have just entered a parallel universe that has only two dimensions! This curious place is inhabited by an intelligent life form, the Flatlanders, who are obsessed with two-dimensional shapes and have a huge vocabulary just for triangles.

Your task is to learn the words that the Flatlanders use for triangles to help us establish contact with these strange beings. It's a pretty difficult task — but we think you're the right person for the job!

### Stage 1: Training

You will see a series of triangles, one by one. Each triangle will be presented with its name in the Flatlander language. The name will also be pronounced by the computer to help you learn it. After every third triangle, you will see one of the previous three triangles again and you must type in its name. This stage is designed to help you learn the language.

### Stage 2: Communication

You and your partner will communicate using the language you learned in Stage 1. When it's your turn to communicate, you'll be presented with a triangle. You will type in the word for this triangle and send it to your partner. Your partner will then see a selection of six triangles and they'll have to figure out which triangle you're talking about. You and your partner will take turns at being the communicator and matcher.

**Important:** you must only communicate using the Flatlander language that you and your partner learned in Stage 1. You must not use English or any other language to communicate with your partner. The supervisor will be monitoring your communications during the experiment.

You will learn a lot of words very quickly during the training stage, and it may be difficult to take it all in. But don't panic! The most important thing is to maintain good relations with the Flatlanders by giving it your best shot. It's okay to guess if you're unsure. Go with your instinct and type in a word that feels right. Your partner may still be able to identify it, even if it's only partially correct.

The pair who communicate most successfully using the Flatlander language will each receive a £20 Amazon voucher.

Good luck!

## Appendix E

### Paper 3, Supplement S2: Geometric measure of triangle dissimilarity

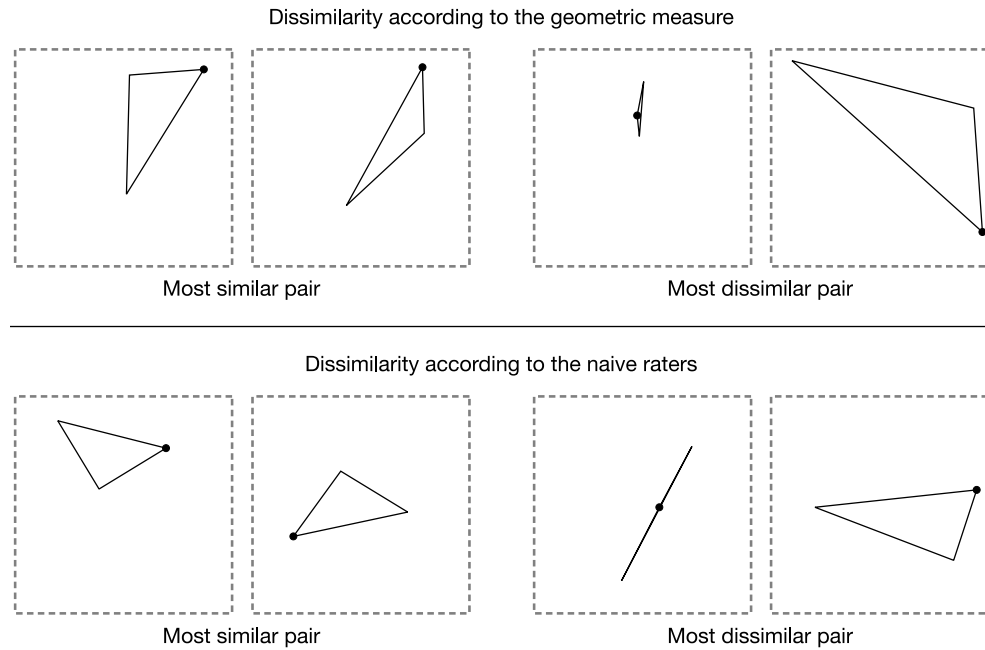
The dissimilarity ratings provided by the naive raters reflect the cognitive perceptions that humans have about triangles. However, the participants in the main experiments may have held different perceptual representations of the triangles from those held by the naive raters: The participants in our main experiments would have had a notion of triangle dissimilarity in the context of using a language to describe or communicate about the triangles. Here I provide an alternative geometric measure of triangle dissimilarity that explicitly considers a range of possible meaning dimensions, some of which may not have been considered important by the naive raters.

Four features were selected that participants could potentially use to conceptualize and communicate about the triangles (see Table E.1). For each of the features, the distance was computed between every pair of triangles in the static set, yielding four

**Table E.1:** Four features and the corresponding distance measures between triangles

Feature	Distance measure
Location	Euclidean distance between centroids
Orientation	Shortest angular distance by orienting spot <sup>a</sup>
Shape	Absolute difference in equilateralness ratio (Equation 2 on page 122)
Size	Absolute difference in centroid size (Note 6 on page 136)

<sup>a</sup>Orientation is defined as the angular coordinate of the orienting spot when the triangle is centred on the origin. The shortest angular distance between two triangles is the shorter of the clockwise or counterclockwise angular distances.

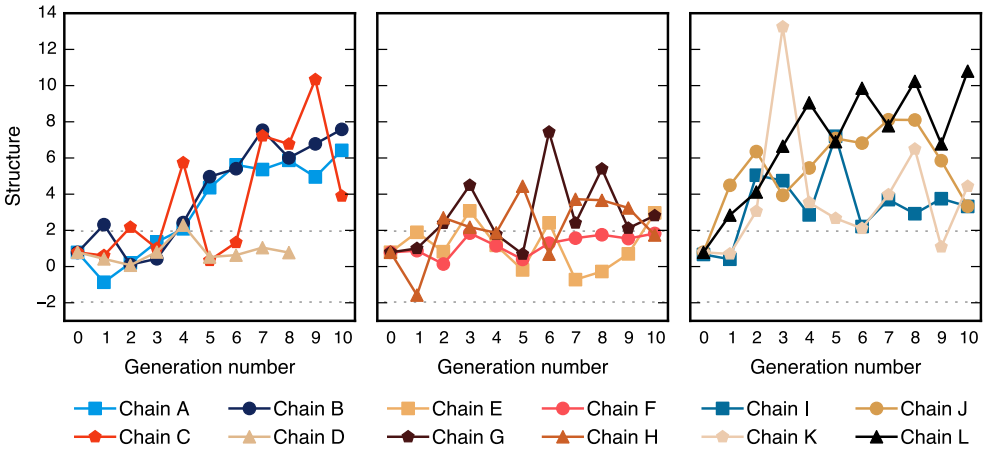


**Figure E.1:** The most similar and most dissimilar pairs of triangles in the static set based on the geometric measure of dissimilarity (top) and the ratings of the naive raters (bottom). Note that, under the geometric measure, all four features are taken into account, while the human raters appear to be ignoring the properties of location and orientation.

distance matrices. The matrices were converted to ranks (in order to remove the distributional effects peculiar to each metric), summed together, and then normalized in the interval  $[0, 1]$ . A pair of triangles that are similar in terms of all four features will have a score close to 0 (with 0 representing identity), while a pair of triangles that are dissimilar in terms of all four features will have a score close to 1. Fig. E.1 shows the most similar and most dissimilar pairs of triangle stimuli based on this geometric measure (top) and based on the ratings from the naive raters (bottom) for comparison.

There was a strong correlation between the scores produced from this geometric approach and the mean normalized dissimilarity ratings provided by the naive raters ( $r = .49$ ,  $n = 1128$ ,  $p < .001$ ; Mantel test). The results for structure using this measure are given in Fig. E.2. The general trends are congruent with the equivalent results produced using the ratings from naive raters. However, the structure scores tend to be lower under the geometric approach, suggesting that it does not fully capture the way in which the triangles are perceived. For this reason, we consider the structure results based on human ratings to be canonical and present this alternative method in support





**Figure E.2:** Levels of structure in Experiment 1, Experiment 2, and Experiment 3 using the geometric measure of triangle dissimilarity rather than human dissimilarity ratings. The results are congruent with those presented in the paper, although the structure scores are generally lower.

of our conclusions.

This geometric measure of triangle dissimilarity described above allows us to isolate particular geometric properties in order to determine which features were being encoded by the participants in the main experiments. To perform this analysis, we correlated the pairwise string dissimilarity scores with all combinations of the four geometric features described in Table E.1 to see which combination would yield the strongest correlation. For four features, there are 15 combinations to consider, yielding 15 different types of language that could potentially arise. These language types are listed in Table E.2 along with reference numbers; for example, a Type 13 language encodes location, shape, and size.

The results of this analysis are given in Table E.3; for all generations, the table gives the type number for the combination of features that resulted in the strongest correlation, along with the Pearson correlation coefficient. The most common types of language to emerge across all experiments were Type 3 (shape; 52% of languages) and Type 10 (shape and size; 17% of languages). The language types with the highest average correlation (across all emergent languages) were Type 3 (shape; mean  $r = .19$ ) and Type 10 (shape and size; mean  $r = .16$ ). These results reveal a clear bias toward encoding the shape and size features of the triangles.

This analysis was also performed with the dissimilarity ratings from the naive raters;

**Table E.2:** List of language types

Type number	Encoded meaning dimensions
1	Location
2	Orientation
3	Shape
4	Size
5	Location, Orientation
6	Location, Shape
7	Location, Size
8	Orientation, Shape
9	Orientation, Size
10	Shape, Size
11	Location, Orientation, Shape
12	Location, Orientation, Size
13	Location, Shape, Size
14	Orientation, Shape, Size
15	Location, Orientation, Shape, Size

**Table E.3:** Encoded meaning dimensions for each emergent language

	1	2	3	4	5	6	7	8	9	10
<b>Experiment 1</b>										
<b>A</b>	1 .06	3 .05	11 .05	14 .07	3 .2	3 .31	3 .45	10 .42	10 .3	3 .46
<b>B</b>	15 .08	3 .05	3 .12	3 .17	3 .4	3 .44	3 .44	3 .52	3 .45	3 .51
<b>C</b>	5 .07	10 .12	3 .07	9 .26	3 .11	3 .06	3 .37	3 .29	3 .47	3 .37
<b>D</b>	14 .05	3 .07	3 .09	13 .17	6 .04	1 .06	–	1 .16	–	–
<b>Experiment 2</b>										
<b>E</b>	11 .08	13 .06	14 .11	10 .05	4 .06	6 .11	3 .07	10 .03	4 .06	7 .19
<b>F</b>	9 .08	4 .04	7 .09	7 .08	7 .03	10 .06	12 .06	10 .08	3 .14	3 .12
<b>G</b>	10 .11	6 .12	3 .23	8 .06	10 .04	10 .31	10 .17	3 .36	6 .1	3 .16
<b>H</b>	4 .0	11 .14	6 .1	10 .08	10 .2	3 .14	3 .21	3 .21	3 .14	6 .1
<b>Experiment 3</b>										
<b>I</b>	10 .08	11 .17	10 .24	3 .18	3 .3	3 .15	3 .33	3 .24	3 .28	3 .25
<b>J</b>	6 .18	3 .33	3 .22	3 .3	3 .36	3 .38	3 .51	3 .37	3 .36	10 .15
<b>K</b>	2 .22	3 .16	3 .45	13 .14	3 .17	3 .17	10 .2	10 .24	3 .14	10 .3
<b>L</b>	1 .14	3 .24	3 .42	3 .49	3 .52	3 .57	3 .41	3 .61	3 .36	10 .48

*Note.* Each cell gives the type number (see Table E.2) for the combination of features that resulted in the strongest correlation, along with the correlation coefficient itself. For example, in Chain A Generation 8, the strongest correlation ( $r = .42$ ) is obtained when the language is assumed to be Type 10, a language that marks shape and size.

the strongest correlations were also with Type 3 (shape;  $r = .71$ ,  $n = 1128$ ,  $p < .001$ ; Mantel test) and Type 10 (shape and size;  $r = .69$ ,  $n = 1128$ ,  $p < .001$ ; Mantel test), suggesting that the naive raters were also rating the dissimilarity between the triangles based primarily on their shape and size features. This is supported by the fact that the dimensions of the MDS solution corresponded approximately to shape and size.



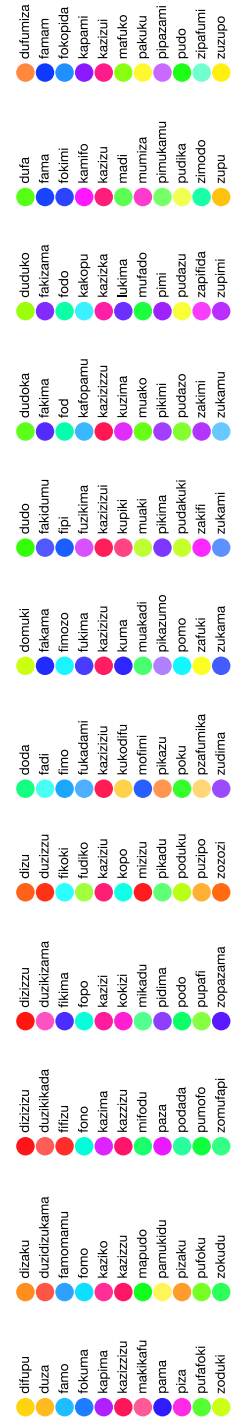
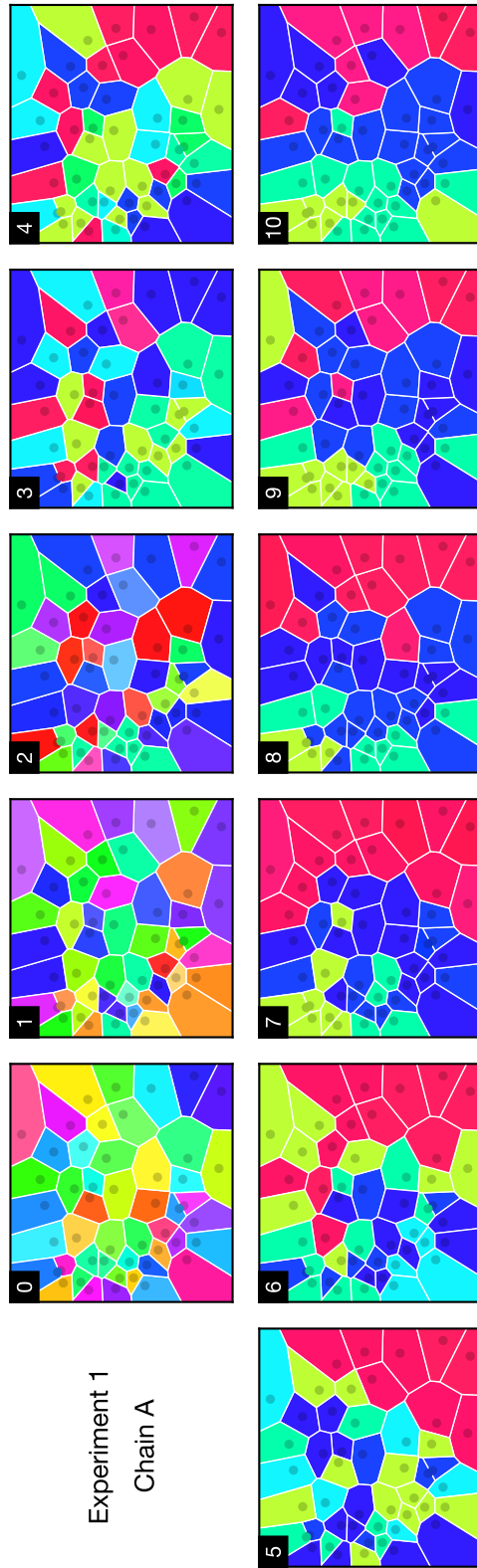
## Appendix F

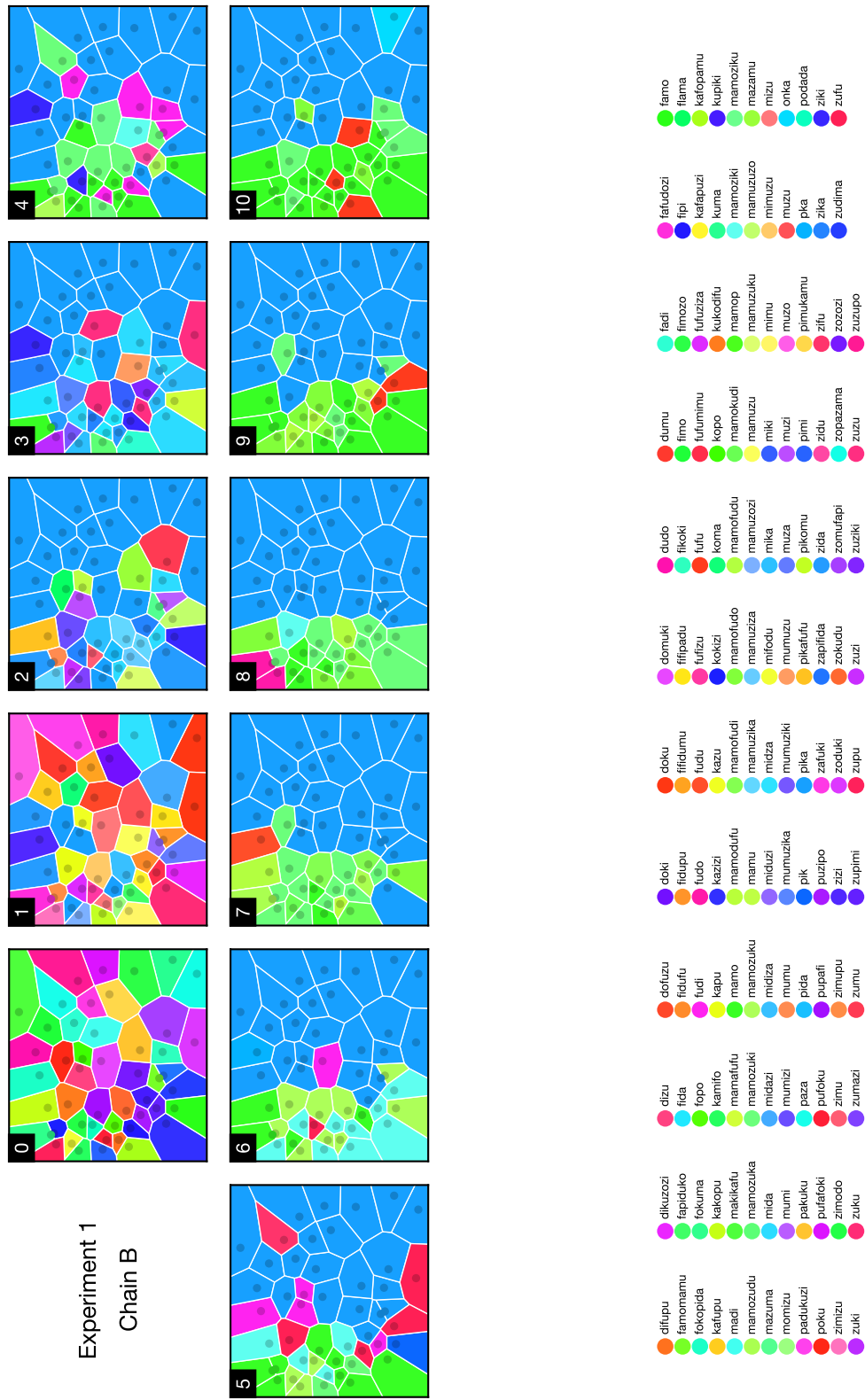
# Paper 3, Supplement S3: MDS plots for all generations in all chains

Each of the subsequent pages depicts the evolution of a single iterated learning chain, from the initial randomly generated labelling at Generation 0 through to Generation 10 (each plot is labelled with its generation number in the top left corner). These plots are generated following the same procedure described in Paper 3 (see pages 120–121). To recap, each plot is a two-dimensional, abstract representation of the space of possible triangles, which was generated by projecting the naive raters' dissimilarity ratings into two dimensions using multidimensional scaling (MDS). Each dark point represents one of the triangles from the static set, and the proximity between two points is indicative of how similar they were judged to be. Roughly, the  $x$ -axis corresponds to the shape of the triangle and the  $y$ -axis corresponds to the size of the triangles.

The colours show how the space was labelled at a given generation. The legend at the bottom of each page gives all unique labels that occurred across the entire chain, and the colours used to represent the labels are selected by fitting an MDS solution to the pairwise Levenshtein edit-distances between them. This makes it possible to track how the languages change over time. At Generation 0 there is no systematic relationship between label similarity (represented by similarity in colour) and meaning similarity (represented by proximity in the space). Over generations, the space begins to become organized into discrete, compact categories: Similar meanings take on similar labels, and as the languages become more structured, they become more reliably transmitted.

Experiment 1  
Chain A

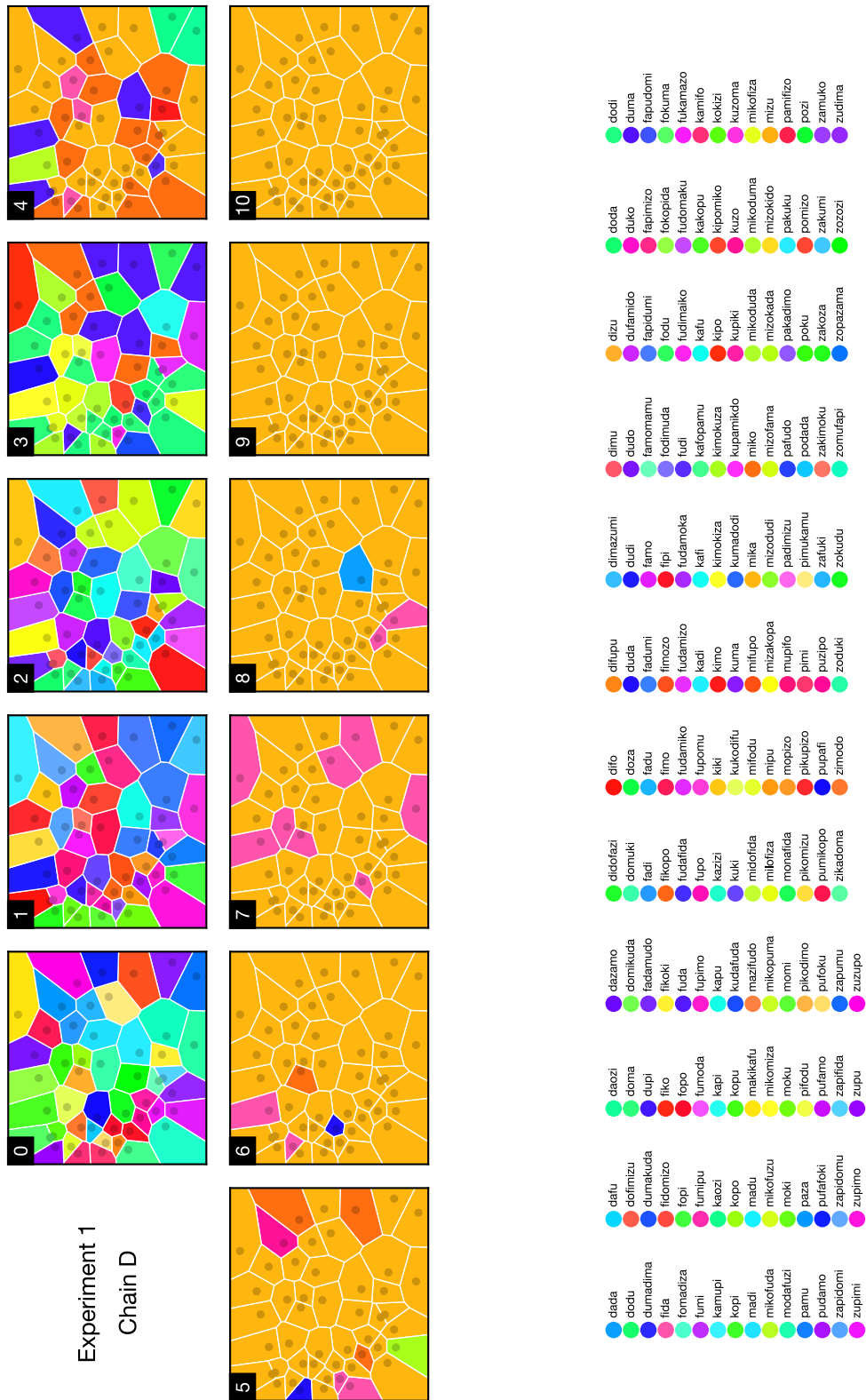




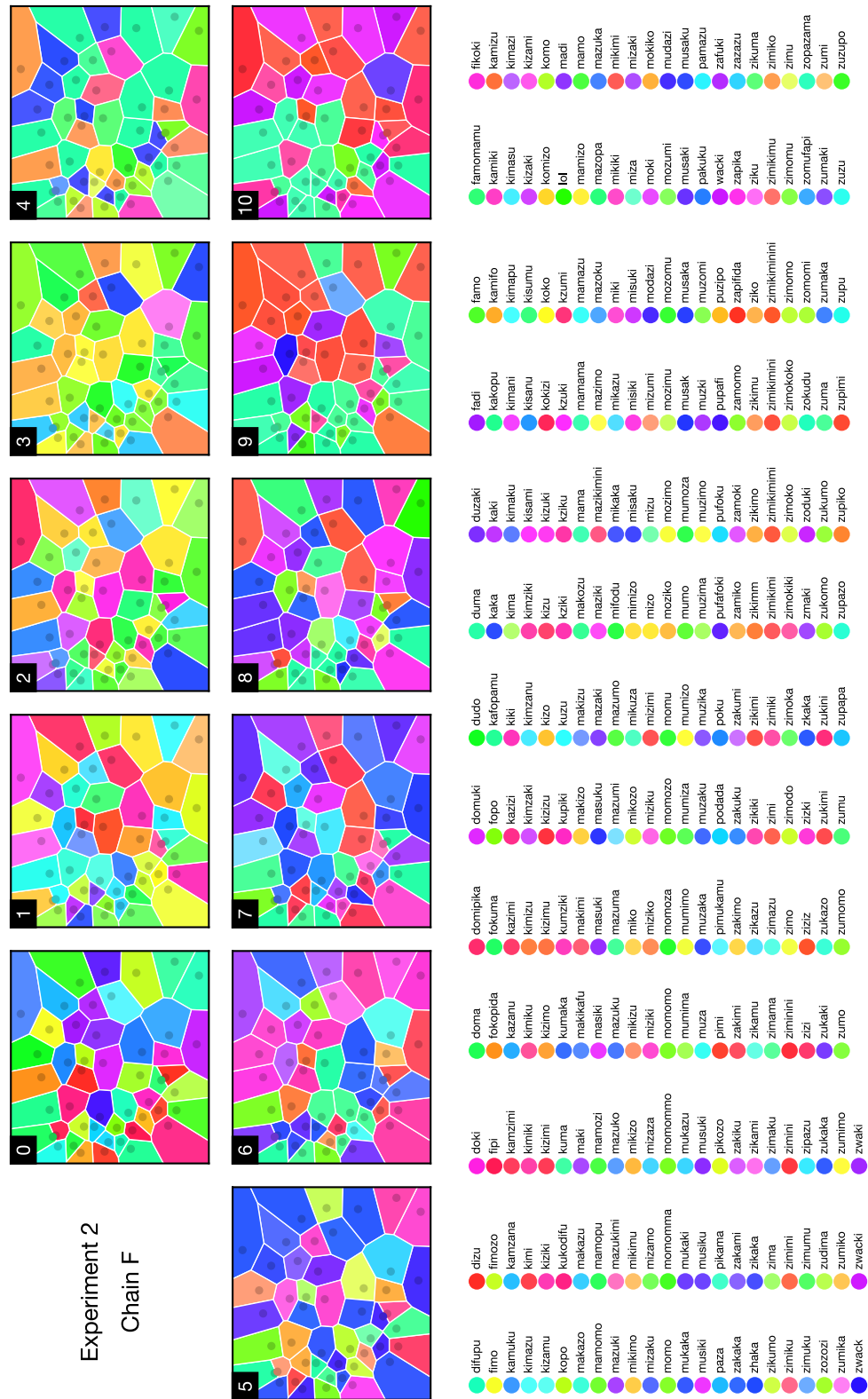
Experiment 1  
Chain C

Figure 1 displays a series of 25 Voronoi diagrams, labeled 0 through 24, arranged in a 5x5 grid. Each diagram represents a spatial partitioning of a domain into regions, likely representing the growth and interaction of different cell lineages or populations over time. The diagrams are color-coded, with colors corresponding to different lineages or populations. The legend on the right lists the names of the lineages or populations, each associated with a specific color. The legend is organized into five columns, each corresponding to a row of diagrams in the grid. The first column of the legend (dafa, dafupu, fadafima, fadafima, fadafima, fadafima, fadafima, fadafima, fadafima, fadafima) corresponds to the first row of diagrams (0-4). The second column (dafa, dafupu, fadafima, fadafima, fadafima, fadafima, fadafima, fadafima, fadafima, fadafima) corresponds to the second row of diagrams (5-9). The third column (dafa, dafupu, fadafima, fadafima, fadafima, fadafima, fadafima, fadafima, fadafima, fadafima) corresponds to the third row of diagrams (10-14). The fourth column (dafa, dafupu, fadafima, fadafima, fadafima, fadafima, fadafima, fadafima, fadafima, fadafima) corresponds to the fourth row of diagrams (15-19). The fifth column (dafa, dafupu, fadafima, fadafima, fadafima, fadafima, fadafima, fadafima, fadafima, fadafima) corresponds to the fifth row of diagrams (20-24).

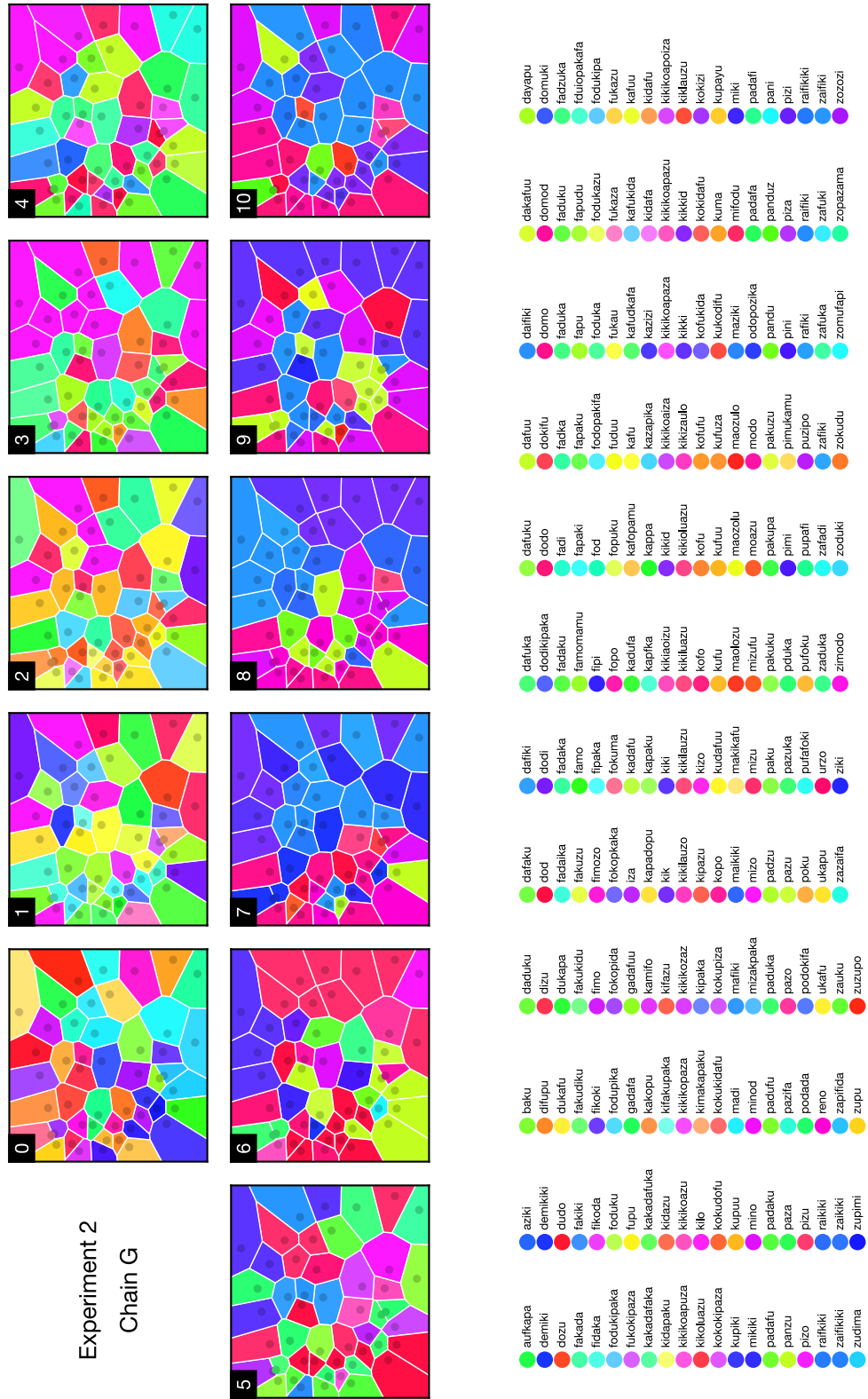




[illegible]

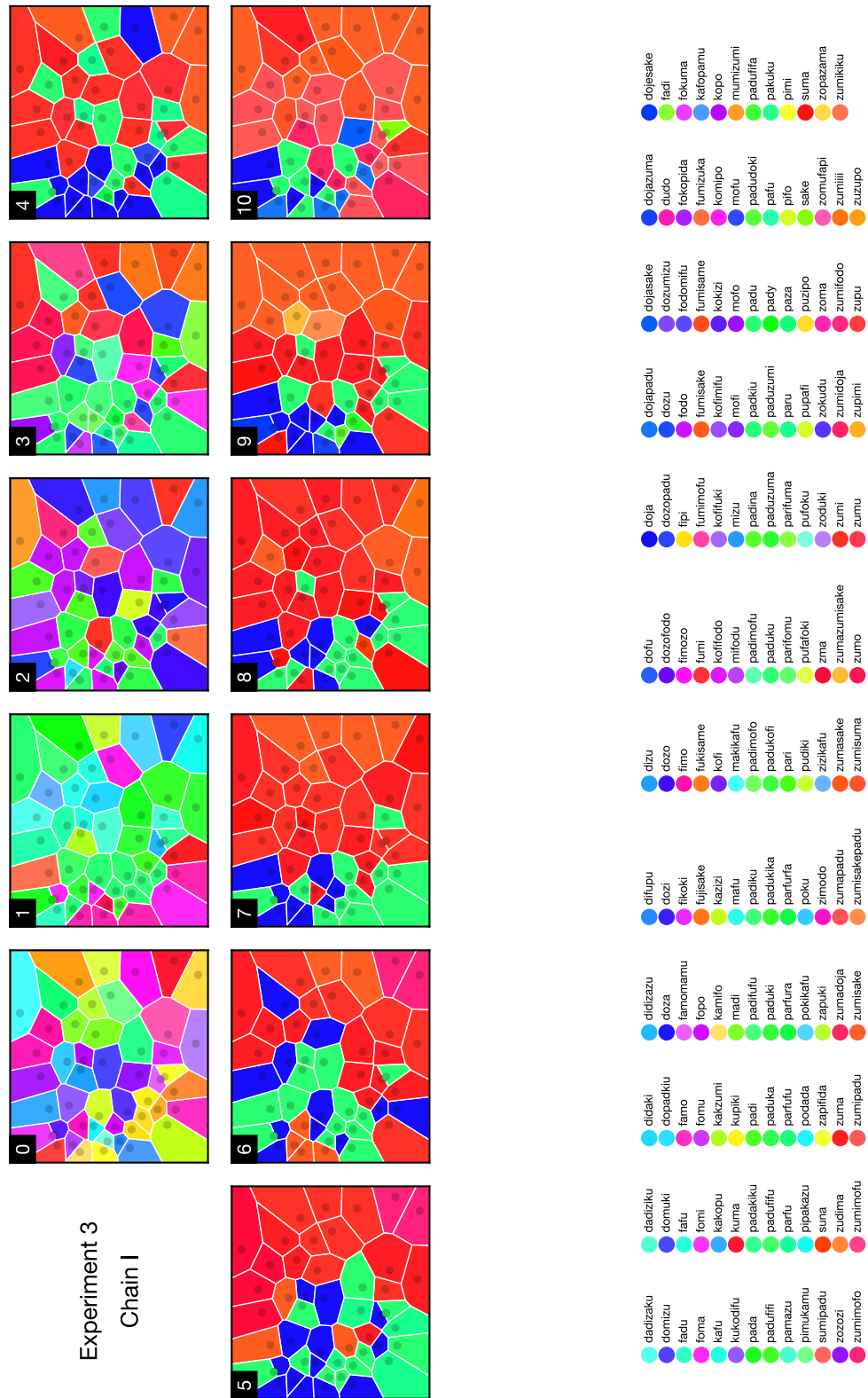


Experiment 2  
Chain G

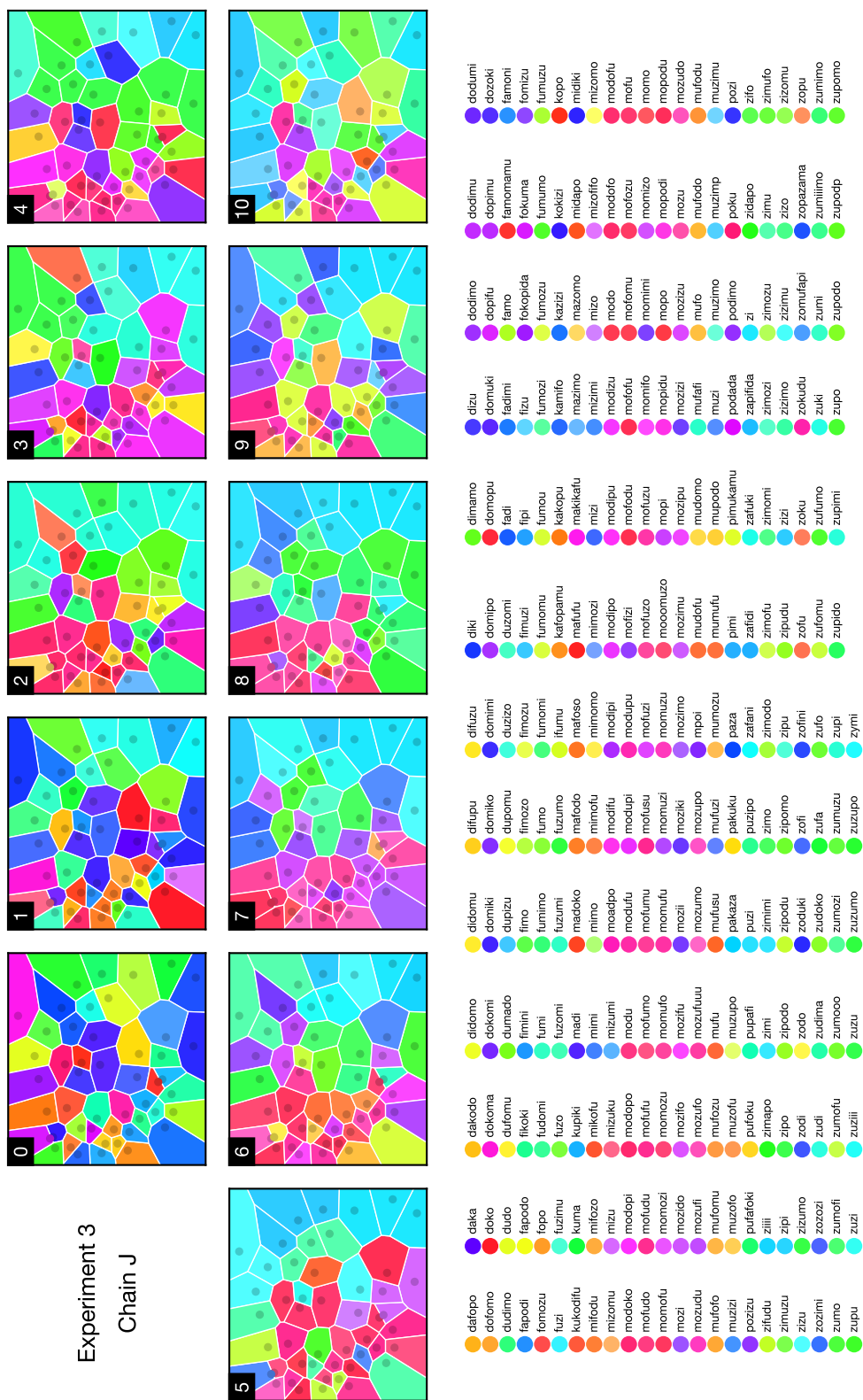


[illegible]

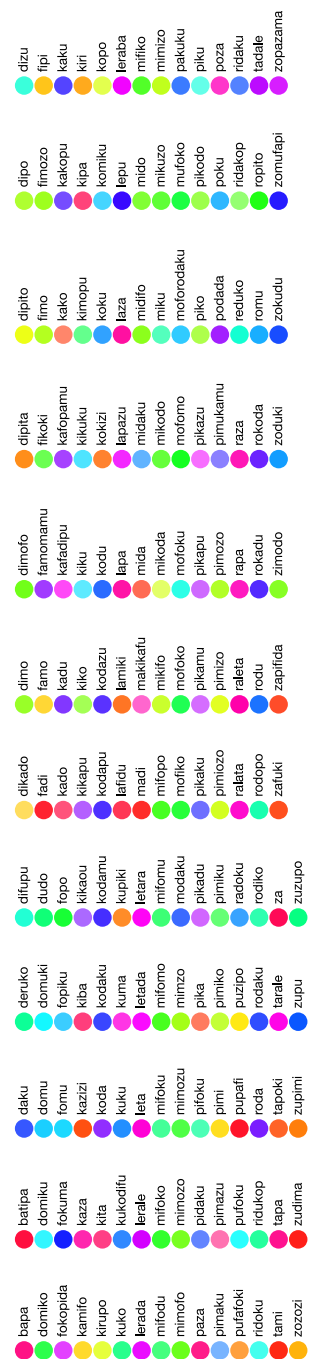
Chain I



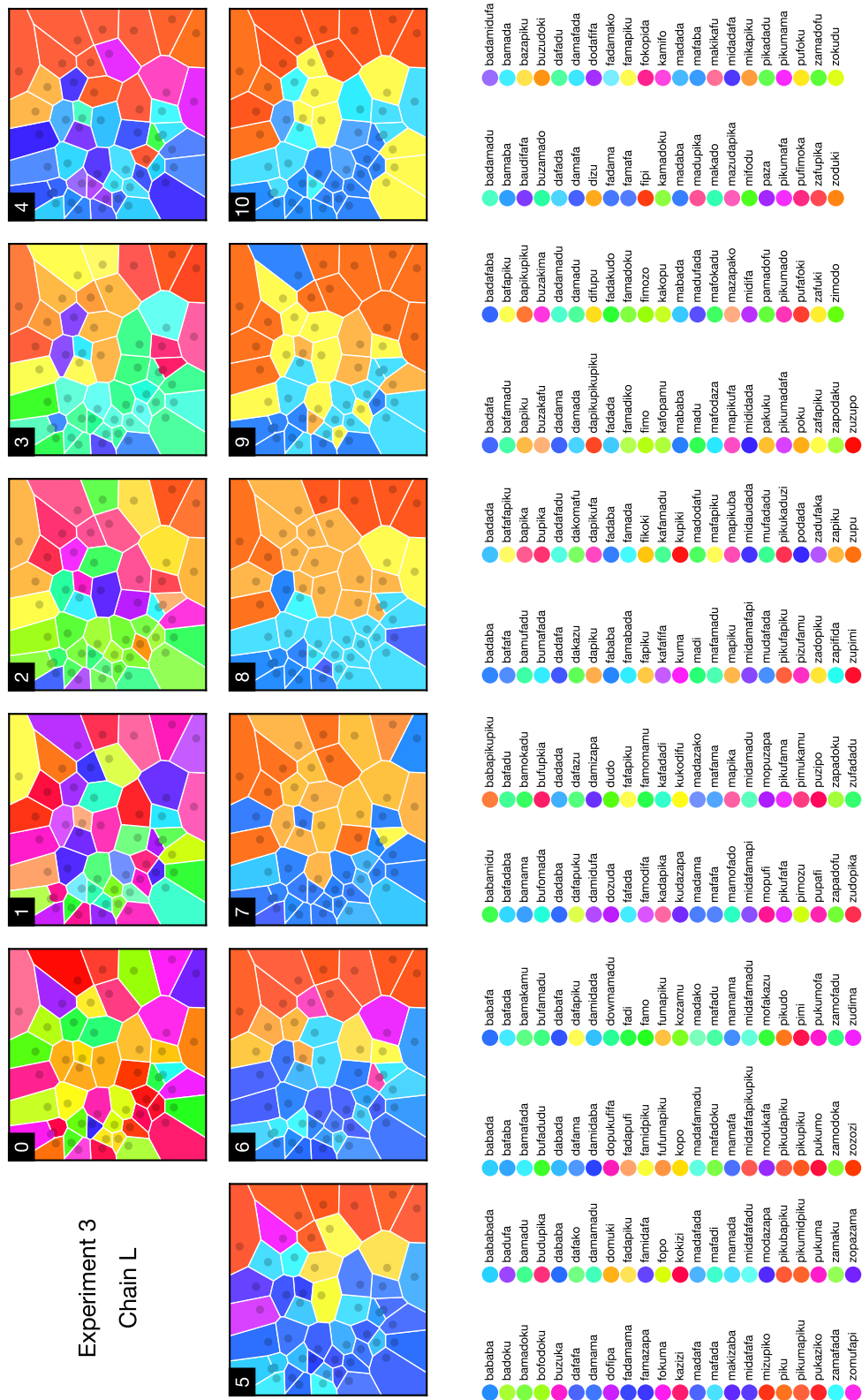




Experiment 3  
Chain K









# References

- Abbott, E. A. (1884). *Flatland: A romance of many dimensions*. London, England: Seeley.
- Ahlner, F., & Zlatev, J. (2010). Cross-modal iconicity: A cognitive semiotic approach to sound symbolism. *Sign Systems Studies*, 38, 298–348.
- Aronoff, M. (1976). *Word formation in generative grammar*. Cambridge, MA: MIT Press.
- Aronoff, M. (2007). In the beginning was the word. *Language*, 83, 803–830. doi:10/dh7j3c
- Ashby, F. G., & Maddox, W. T. (1990). Integrating information from separable psychological dimensions. *Journal of Experimental Psychology: Human Perception and Performance*, 16, 598–612. doi:10/crz4hz
- Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*, 56, 149–178. doi:10/fmnwkc
- Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge, England: Cambridge University Press. doi:10/fhdtmk
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67, 1–48. doi:10/gcrnkw
- Beckner, C., Pierrehumbert, J. B., & Hay, J. (2017). The emergence of linguistic structure in an online iterated learning task. *Journal of Language Evolution*, 2, 160–176. doi:10/gfr9mq
- Bergen, B. K. (2004). The psychological reality of phonaesthemes. *Language*, 80, 290–311. doi:10/dztbz5
- Bikkers, A. V. W. (1869). *Darwinism tested by the science of language*. London, England: John Camden Hotten.

- Bookstein, F. L. (1991). *Morphometric tools for landmark data: Geometry and biology*. Cambridge, England: Cambridge University Press. doi:10/cf3kjk
- Borg, I., & Groenen, P. J. F. (2005). *Modern multidimensional scaling: Theory and applications* (2nd ed.). New York, NY: Springer-Verlag. doi:10/d8hgdv
- Bowyer, A. (1981). Computing Dirichlet tessellations. *The Computer Journal*, 24, 162–166. doi:10/d47xbw
- Brighton, H. (2002). Compositional syntax from cultural transmission. *Artificial Life*, 8, 25–54. doi:10/ctf97r
- Brighton, H., Smith, K., & Kirby, S. (2005). Language as an evolutionary system. *Physics of Life Reviews*, 2, 177–226. doi:10/dqbhzz
- Caldwell, C. A., & Millen, A. E. (2008). Experimental models for testing hypotheses about cumulative cultural evolution. *Evolution and Human Behavior*, 29, 165–171. doi:10/fhqgrd
- Canini, K. R., Griffiths, T. L., Vanpaemel, W., & Kalish, M. L. (2014). Revealing human inductive biases for category learning by simulating cultural transmission. *Psychonomic Bulletin & Review*, 21, 785–793. doi:10/f55qdm
- Carr, J. W. (2013). *The cumulative cultural evolution of category structure in an infinite meaning space* (Unpublished master's thesis). University of Edinburgh, Scotland.
- Carr, J. W., & Smith, K. (2016). Modeling language transmission. In V. A. Weekes-Shackelford & T. K. Weekes-Shackelford (Eds.), *Encyclopedia of evolutionary psychological science*. Springer. doi:10/gfr9hb
- 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, 892–923. doi:10/f98sr2
- Carstensen, A., Xu, J., Smith, C. T., & Regier, T. (2015). Language evolution in the lab tends toward informative communication. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 303–308). Austin, TX: Cognitive Science Society.
- Chaitin, G. J. (1969). On the simplicity and speed of programs for computing infinite sets of natural numbers. *Journal of the Association for Computing Machinery*, 16, 407–422. doi:10/dnzzth
- Chater, N., Clark, A., Goldsmith, J. A., & Perfors, A. (2015). *Empiricism and language*

- learnability*. Oxford, England: Oxford University Press. doi:10/gfr9g9
- Chater, N., & Vitányi, P. M. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences*, 7, 19–22. doi:10/bbjn6m
- Chater, N., & Vitányi, P. M. (2007). ‘Ideal learning’ of natural language: Positive results about learning from positive evidence. *Journal of Mathematical Psychology*, 51, 135–163. doi:10/bkm7cv
- Chekaf, M., Gauvrit, N., Guida, A., & Mathy, F. (2018). Compression in working memory and its relationship with fluid intelligence. *Cognitive Science*, 42, 904–922. doi:10/gdqzwf
- Cheung, H.-n. S. (1990). Terms of address in Cantonese. *Journal of Chinese Linguistics*, 18, 1–43.
- Chomsky, N. (1980). *Rules and representations*. Oxford, England: Blackwell.
- Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, 31, 489–558. doi:10/bpvn67
- Claidière, N., Smith, K., Kirby, S., & Fagot, J. (2014). Cultural evolution of systematically structured behaviour in a non-human primate. *Proceedings of the Royal Society B: Biological Sciences*, 281, 1–9. doi:10/gfr9gt
- Contreras Kallens, P. A., Dale, R., & Smaldino, P. E. (2018). Cultural evolution of categorization. *Cognitive Systems Research*, 52, 765–774. doi:10/gfq3xw
- Cornish, H. (2011). *Language adapts: Exploring the cultural dynamics of iterated learning* (Unpublished doctoral dissertation). University of Edinburgh, Scotland.
- Culbertson, J., & Kirby, S. (2016). Simplicity and specificity in language: Domain-general biases have domain-specific effects. *Frontiers in Psychology*, 6, 1–11. doi:10/gccx6k
- Cuskley, C., & Kirby, S. (2013). Synesthesia, cross-modality, and language evolution. In J. Simner & E. M. Hubbard (Eds.), *The Oxford handbook of synesthesia* (pp. 869–899). Oxford, England: Oxford University Press. doi:10/gfr9mw
- Darwin, C. (1871). *The descent of man, and selection in relation to sex*. London, England: John Murray.
- de Boer, B., & Thompson, B. (2018). Biology-culture co-evolution in finite populations. *Scientific Reports*, 8, 1209. doi:10/czmv
- del Giudice, A. (2012). The emergence of duality of patterning through iterated learn-

- ing: Precursors to phonology in a visual lexicon. *Language and Cognition*, 4, 381–418. doi:10/gfr9g8
- de Saussure, F. (1959). *Course in general linguistics*. New York, NY: Philosophical Library.
- Dingemanse, M. (2013, May 28). Evolving words: Darwin on Müller on Schleicher [Blog post]. *Diversity Linguistics Comment*. Retrieved from <https://dlc.hypotheses.org/399>
- Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, iconicity, and systematicity in language. *Trends in Cognitive Sciences*, 19, 603–615. doi:10/f7t2x6
- Dobzhansky, T. (1973). Nothing in biology makes sense except in the light of evolution. *The American Biology Teacher*, 35, 125–129. doi:10/gcjqmr
- Douven, I., & Gärdenfors, P. (in press). What are natural concepts? A design perspective. *Mind & Language*.
- Eppstein, D. (2010). Graph-theoretic solutions to computational geometry problems. In C. Paul & M. Habib (Eds.), *Graph-theoretic concepts in computer science* (pp. 1–16). Berlin, Germany: Springer. doi:10/bc8q6h
- Erickson, M. A., & Kruschke, J. K. (2002). Rule-based extrapolation in perceptual categorization. *Psychonomic Bulletin & Review*, 9, 160–168. doi:10/d3tw86
- Esper, E. A. (1966). Social transmission of an artificial language. *Language*, 42, 575–580. doi:10/bmbb98
- Fass, D., & Feldman, J. (2002). Categorization under complexity: A unified MDL account of human learning of regular and irregular categories. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems 15* (pp. 35–42). Cambridge, MA: MIT Press.
- Fedzechkina, M., Jaeger, T. F., & Newport, E. L. (2012). Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences of the USA*, 109, 17897–17902. doi:10/f4dsqk
- Fehér, O., Wang, H., Saar, S., Mitra, P. P., & Tchernichovski, O. (2009). *De novo* establishment of wild-type song culture in the zebra finch. *Nature*, 459, 564–568. doi:10/dbqh7w
- Feldman, J. (2000). Minimization of Boolean complexity in human concept learning.

- Nature*, 407, 630–633. doi:10/cwndrg
- Feldman, J. (2016). The simplicity principle in perception and cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 7, 330–340. doi:10/f824bw
- Ferdinand, V. (2015). *Inductive evolution: Cognition, culture, and regularity in language* (Unpublished doctoral dissertation). University of Edinburgh, Scotland.
- Ferdinand, V., Kirby, S., & Smith, K. (2019). The cognitive roots of regularization in language. *Cognition*, 184, 53–68. doi:10/czgf
- Ferrari, L., Sankar, P. V., & Sklansky, J. (1984). Minimal rectangular partitions of digitized blobs. *Computer Vision, Graphics, and Image Processing*, 28, 58–71. doi:10/cp4zzq
- Ferrer i Cancho, R., & Solé, R. V. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences of the USA*, 100, 788–791. doi:10/c7r62f
- Fitch, W. T. (2010). *The evolution of language*. Cambridge, England: Cambridge University Press.
- Flaherty, M., & Kirby, S. (2008). Iterated language learning in children. In A. D. M. Smith, K. Smith, & R. Ferrer i Cancho (Eds.), *The evolution of language: Proceedings of the 7th international conference* (pp. 425–426). Singapore: World Scientific.
- Frank, M. C., Goodman, N., Lai, P., & Tenenbaum, J. B. (2009). Informative communication in word production and word learning. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 1228–1233). Austin, TX: Cognitive Science Society.
- Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336, 998–998. doi:10.1126/science.1218633
- Frank, M. C., & Goodman, N. D. (2014). Inferring word meanings by assuming that speakers are informative. *Cognitive Psychology*, 75, 80–96. doi:10/f6vfr6
- Freeman, H. (1961). On the encoding of arbitrary geometric configurations. *IEEE Transactions on Electronic Computers*, EC-10, 260–268. doi:10/brg835
- Galantucci, B. (2005). An experimental study of the emergence of human communication systems. *Cognitive Science*, 29, 737–767. doi:10/cx9vp2
- Gärdenfors, P. (2000). *Conceptual spaces: The geometry of thought*. Cambridge, MA:

- MIT Press.
- Gärdenfors, P. (2014). *The geometry of meaning: Semantics based on conceptual spaces*. Cambridge, MA: MIT Press.
- Garrod, S., Fay, N., Lee, J., Oberlander, J., & MacLeod, T. (2007). Foundations of representation: Where might graphical symbol systems come from? *Cognitive Science*, 31, 961–987. doi:10/b4n2k2
- Gibson, E., Futrell, R., Jara-Ettinger, J., Mahowald, K., Bergen, L., Ratnasingam, S., ... Conway, B. R. (2017). Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences of the USA*, 114, 10785–10790. doi:10/cc8g
- Giordano, B. L., Guastavino, C., Murphy, E., Ogg, M., Smith, B. K., & McAdams, S. (2011). Comparison of methods for collecting and modeling dissimilarity data: Applications to complex sound stimuli. *Multivariate Behavioral Research*, 46, 779–811. doi:10/dwpt9b
- Goudbeek, M., Swingle, D., & Smits, R. (2009). Supervised and unsupervised learning of multidimensional acoustic categories. *Journal of Experimental Psychology: Human Perception and Performance*, 35, 1913–1933. doi:10/ftqfkv
- Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science*, 31, 441–480. doi:10/dtbng4
- Grünwald, P. D. (2007). *The minimum description length principle*. Cambridge, MA: MIT Press.
- 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, 1569–1579. doi:10/dx5cg8
- Head, T., MechCoder, Louppe, G., Shcherbatyi, I., fcharras, Vinícius, Z., ... Fabisch, A. (2018). *scikit-optimize: v0.5.1*. doi:10/gfr9mx
- Hockett, C. F. (1960). The origin of speech. *Scientific American*, 203, 88–96.
- Hofstadter, D. R. (2001). Epilogue: Analogy as the core of cognition. In D. Gentner, K. J. Holyoak, & B. N. Kokinov (Eds.), *The analogical mind: Perspectives from cognitive science* (pp. 499–538). Cambridge, MA: MIT Press.
- Hopcroft, J. E., & Karp, R. M. (1973). An  $n^{5/2}$  algorithm for maximum matchings in bipartite graphs. *SIAM Journal on Computing*, 2, 225–231. doi:10/bbjsgb
- Horner, V., Whiten, A., Flynn, E., & de Waal, F. B. M. (2006). Faithful replication



- of foraging techniques along cultural transmission chains by chimpanzees and children. *Proceedings of the National Academy of Sciences of the USA*, 103, 13878–13883. doi:10/fcnc7f
- Hurford, J. R. (1989). Biological evolution of the Saussurean sign as a component of the language acquisition device. *Lingua*, 77, 187–222. doi:10/bgrjnz
- Hutchins, S. S. (1998). *The psychological reality, variability, and compositionality of English phonesthemes* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 9901857)
- Jäger, G. (2010). Natural color categories are convex sets. In M. Aloni, H. Bastiaanse, T. Jäger, & K. Schulz (Eds.), *Logic, language and meaning* (pp. 11–20). Berlin, Germany: Springer Berlin Heidelberg. doi:10/c727xg
- Jäger, G., & van Rooij, R. (2007). Language structure: Psychological and social constraints. *Synthese*, 159, 99–130. doi:10/b46ch2
- Johansson, N., Carr, J. W., & Kirby, S. (in prep.). How to create sound symbolism: Direct transmission of sound strings produces sound–meaning associations.
- 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. doi:10/gbkfnp
- Kay, P., Berlin, B., Maffi, L., Merrifield, W. R., & Cook, R. (2009). *The world color survey*. Stanford, CA: Center for the Study of Language and Information.
- Kemp, C. (2012). Exploring the conceptual universe. *Psychological Review*, 119, 685–722. doi:10/f4b8qk
- Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, 336, 1049–1054. doi:10/f3zjdp
- Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics*, 4, 109–128. doi:10/gfpnnj
- Kempe, V., Gauvrit, N., & Forsyth, D. (2015). Structure emerges faster during cultural transmission in children than in adults. *Cognition*, 136, 247–254. doi:10/f626vj
- Khetarpal, N., Neveu, G., Majid, A., Michael, L., & Regier, T. (2013). Spatial terms across languages support near-optimal communication: Evidence from Peruvian Amazonia, and computational analyses. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive*

- Science Society* (pp. 764–769). Austin, TX: Cognitive Science Society.
- Kirby, S. (1999). *Function, selection, and innateness: The emergence of language universals*. Oxford, England: Oxford University Press.
- Kirby, S. (2002). Learning, bottlenecks and the evolution of recursive syntax. In E. J. Briscoe (Ed.), *Linguistic evolution through language acquisition: Formal and computational models* (pp. 173–203). Cambridge, England: Cambridge University Press. doi:10/bz5gkw
- Kirby, S. (2007). The evolution of meaning-space structure through iterated learning. In C. Lyon, C. L. Nehaniv, & A. Cangelosi (Eds.), *Emergence of communication and language* (pp. 253–267). London, England: Springer-Verlag. doi:10/dzst6z
- Kirby, S. (2014). The evolution of Evolang (Simon Kirby). In E. A. Cartmill, S. G. Roberts, H. Lyn, & H. Cornish (Eds.), *The evolution of language: Proceedings of the 10th international conference* (p. 15). Singapore: World Scientific. doi:10/gfr9hc
- 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 of the USA*, 105, 10681–10686. doi:10/bm7k2r
- Kirby, S., Griffiths, T. L., & Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, 28, 108–114. doi:10/f6mpz3
- Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141, 87–102. doi:10/f7jcnn
- Köhler, W. (1929). *Gestalt psychology*. New York, NY: Liveright.
- Kolmogorov, A. N. (1965). Три подхода к определению понятия «количество информации» [Three approaches to the definition of the concept “quantity of information”]. *Проблемы Передачи Информации*, 1, 3–11.
- Kovic, V., Plunkett, K., & Westermann, G. (2010). The shape of words in the brain. *Cognition*, 114, 19–28. doi:10/bgg3w8
- Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30, 61–70. doi:10/b6v5kt

- Kwon, N., & Round, E. R. (2015). Phonaesthemes in morphological theory. *Morphology*, 25, 1–27. doi:10/gfr9mr
- Labov, W. (1973). The boundaries of words and their meanings. In C.-J. N. Bailey & R. W. Shuy (Eds.), *New ways of analyzing variation in English* (pp. 340–373). Washington, DC: Georgetown University Press.
- Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago, IL: The University of Chicago Press.
- Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, 3, 299–321. doi:10/bhwkff
- Laskowski, C. (2008). The emergence of a lexicon by prototype-categorising agents in a structured infinite world. In A. D. M. Smith, K. Smith, & R. Ferrer i Cancho (Eds.), *The evolution of language: Proceedings of the 7th international conference* (pp. 195–202). Singapore: World Scientific. doi:10/fmpnx7
- Lespinats, S., & Fertil, B. (2011). ColorPhylo: A color code to accurately display taxonomic classifications. *Evolutionary Bioinformatics*, 7, 257–270. doi:10/df9fwr
- Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10, 707–710.
- Levinson, S. C. (2012). Kinship and human thought. *Science*, 336, 988–989. doi:10/gfr9gv
- Li, M., & Vitányi, P. M. (2008). *An introduction to Kolmogorov complexity and its applications*. New York, NY: Springer. doi:10/fkccvf
- Lipski Jr, W., Lodi, E., Luccio, F., Mugnai, C., & Pagli, L. (1979). On two-dimensional data organization II. *Fundamenta Informaticae*, 2, 245–260.
- Little, H., Eryilmaz, K., & de Boer, B. (2017). Signal dimensionality and the emergence of combinatorial structure. *Cognition*, 168, 1–15. doi:10/gbz6m4
- Lockwood, G., & Dingemanse, M. (2015). Iconicity in the lab: A review of behavioral, developmental, and neuroimaging research into sound-symbolism. *Frontiers in Psychology*, 6, 1–14. doi:10/gfr9ms
- Lupyan, G. (2017). The paradox of the universal triangle: Concepts, language, and prototypes. *The Quarterly Journal of Experimental Psychology*, 70, 389–412. doi:10/gcx83s
- Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not just for talking.

- Psychological Science*, 18, 1077–1083. doi:10/bx5ct5
- Malt, B. C., Sloman, S. A., & Gennari, S. P. (2003). Universality and language specificity in object naming. *Journal of Memory and Language*, 49, 20–42. doi:10/dk84t7
- Malt, B. C., Sloman, S. A., Gennari, S. P., Shi, M., & Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language*, 40, 230–262. doi:10/d3vjk6
- Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27, 209–220.
- Martinet, A. (1952). Function, structure, and sound change. *Word*, 8, 1–32. doi:10/gfr9gw
- Mathy, F., & Feldman, J. (2012). What's magic about magic numbers? Chunking and data compression in short-term memory. *Cognition*, 122, 346–362. doi:10/ds3s2z
- Maurer, D., Pathman, T., & Mondloch, C. J. (2006). The shape of boubas: Sound-shape correspondences in toddlers and adults. *Developmental Science*, 9, 316–322. doi:10/cpx5mk
- Meilă, M. (2007). Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, 98, 873–895. doi:10/ftwhz3
- 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, 325–347. doi:10/cnkx88
- Monaghan, P., Shillcock, R. C., Christiansen, M. H., & Kirby, S. (2014). How arbitrary is language? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369, 1–12. doi:10/f6p6g7
- Moravcsik, E. (1978). Reduplicative constructions. In J. H. Greenberg (Ed.), *Universals of human language: Word structure* (Vol. 3, pp. 297–334). Stanford, CA: Stanford University Press.
- Moreton, E., Pater, J., & Pertsova, K. (2015). Phonological concept learning. *Cognitive Science*, 41, 4–69. doi:10/f9vs9g
- Motamedi, Y., Schouwstra, M., Smith, K., Culbertson, J., & Kirby, S. (under review). Evolving artificial sign languages in the lab: From improvised gesture to systematic sign. Retrieved from <https://psyarxiv.com/be7qy/>

- Müller, M. (1870). *Darwinism tested by the science of language*. Translated from the German of Professor August Schleicher. *Nature*, 1, 256–259. doi:10/b3k76q
- Murdock, G. P. (1970). Kin term patterns and their distribution. *Ethnology*, 9, 165–207. doi:10/ds5jgg
- Murphy, G. L. (2004). *The big book of concepts*. Cambridge, MA: MIT Press.
- Newton, I. (1729). *The mathematical principles of natural philosophy*. London, England: Benjamin Motte.
- Nielsen, A., & Rendall, D. (2012). The source and magnitude of sound-symbolic biases in processing artificial word material and their implications for language learning and transmission. *Language and Cognition*, 4, 115–125. doi:10/gfr9mt
- Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, 115, 39–57. doi:10/bgj3w
- Nosofsky, R. M., Palmeri, T. J., & Mckinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, 101, 53–79. doi:10/c7j47d
- Nowak, M. A., & Krakauer, D. C. (1999). The evolution of language. *Proceedings of the National Academy of Sciences of the USA*, 96, 8028–8033. doi:10/brb9w4
- Nuckolls, J. B. (1999). The case for sound symbolism. *Annual Review of Anthropology*, 28, 225–252. doi:10/fb6qkp
- Nygaard, L. C., Cook, A. E., & Namy, L. L. (2009). Sound to meaning correspondences facilitate word learning. *Cognition*, 112, 181–186. doi:10/dj95vg
- Ohtsuki, T. (1982). Minimum dissection of rectilinear regions. In *Proceedings of the IEEE International Conference on Circuits and Systems* (pp. 1210–1213).
- Oliphant, M. (1996). The dilemma of Saussurean communication. *BioSystems*, 37, 31–38. doi:10/bpzn6k
- Page, E. (1963). Ordered hypotheses for multiple treatments: A significance test for linear ranks. *Journal of the American Statistical Association*, 58, 216–230. doi:10/gfr9gx
- Parault, S., & Schwanenflugel, P. (2006). Sound-symbolism: A piece in the puzzle of word learning. *Journal of Psycholinguistic Research*, 35, 329–351. doi:10/dzwzss
- Perfors, A., & Navarro, D. J. (2014). Language evolution can be shaped by the structure of the world. *Cognitive Science*, 38, 775–793. doi:10/f56xfq

- Perlman, M., Dale, R., & Lupyan, G. (2015). Iconicity can ground the creation of vocal symbols. *Royal Society Open Science*, 2, 1–16. doi:10/czmr
- Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences of the USA*, 108, 3526–3529. doi:10/cdj3vc
- Pinker, S., & Bloom, P. (1990). Natural language and natural selection. *Behavioral and Brain Sciences*, 13, 707–784. doi:10/fbrnvn
- Plotkin, J. B., & Nowak, M. A. (2000). Language evolution and information theory. *Journal of Theoretical Biology*, 205, 147–159. doi:10/fqmq3n
- Ravignani, A., Delgado, T., & Kirby, S. (2016). Musical evolution in the lab exhibits rhythmic universals. *Nature Human Behaviour*, 1, 1–7. doi:10/gfr9gz
- 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. doi:10/gfhrzq
- Raviv, L., Meyer, A., & Lev-Ari, S. (2018). Compositional structure can emerge without generational transmission. *Cognition*, 182, 151–164. doi:10/gfr9mv
- Regier, T. (1998). Reduplication and the arbitrariness of the sign. In M. Gernsbacher & S. Derry (Eds.), *Proceedings of the 20th Annual Conference of the Cognitive Science Society* (pp. 887–892). Mahwah, NJ: Lawrence Erlbaum Associates.
- Regier, T., Carstensen, A., & Kemp, C. (2016). Languages support efficient communication about the environment: Words for snow revisited. *PLOS ONE*, 11, e0151138. doi:10/f8w4sx
- Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences of the USA*, 104, 1436–1441. doi:10/cvhcnd
- Regier, T., Kemp, C., & Kay, P. (2015). Word meanings across languages support efficient communication. In B. MacWhinney & W. O'Grady (Eds.), *The handbook of language emergence* (pp. 237–263). Hoboken, NJ: John Wiley & Sons. doi:10/gfr9hd
- Richerson, P. J., & Boyd, R. (2005). *Not by genes alone: How culture transformed human evolution*. Chicago, IL: University of Chicago Press.
- Richie, R. (2016). Functionalism in the lexicon: Where is it, and how did it get there?

- The Mental Lexicon*, 11, 429–466. doi:10/gfr9g2
- Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14, 465–471. doi:10/c4gt7t
- Rissanen, J. (1989). *Stochastic complexity in statistical inquiry*. Singapore: World Scientific.
- Rosch, E. (1973). Natural categories. *Cognitive Psychology*, 4, 328–350. doi:10/cj6qyb
- Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–48). Hillsdale, NJ: Lawrence Erlbaum.
- Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7, 573–605. doi:10/d7d639
- Saramago, J. (2011). *Cain* (M. Jull Costa, Trans.). London, England: Harvill Secker.
- Schleicher, A. (1863). *Die Darwinsche Theorie und die Sprachwissenschaft* [Darwinian theory and the language sciences]. Weimar, Germany: Hermann Böhlau.
- Scott-Phillips, T. C., & Kirby, S. (2010). Language evolution in the laboratory. *Trends in Cognitive Sciences*, 14, 411–417. doi:10/dfjm9w
- Scott-Phillips, T. C., Kirby, S., & Ritchie, G. R. S. (2009). Signalling signalhood and the emergence of communication. *Cognition*, 113, 226–233. doi:10/fs65qm
- Selten, R., & Warglien, M. (2007). The emergence of simple languages in an experimental coordination game. *Proceedings of the National Academy of Sciences of the USA*, 104, 7361–7366. doi:10/bh7q7d
- Shamir, O., Sabato, S., & Tishby, N. (2010). Learning and generalization with the information bottleneck. *Theoretical Computer Science*, 411, 2696–2711. doi:10/dc89wx
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27, 379–423.
- Shannon, C. E. (1956). The bandwagon. *IRE Transactions on Information Theory*, 2, 3. doi:10/fqvwvh
- Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana, IL: University of Illinois Press.
- Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, 1, 54–87. doi:10/fs3zxm
- Shepard, R. N. (1987). Toward a universal law of generalization for psychological sci-

- ence. *Science*, 237, 1317–1323. doi:10/cwsqdz
- Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, 75, 1–42. doi:10/cp7tv3
- Shillcock, R. C., Kirby, S., McDonald, S., & Brew, C. (2001). Filled pauses and their status in the mental lexicon. In *Proceedings of the 2001 conference of disfluency in spontaneous speech* (pp. 53–56).
- Silvey, C. (2014). *The communicative emergence and cultural evolution of word meanings* (Unpublished doctoral dissertation). University of Edinburgh, Scotland.
- Silvey, C., Kirby, S., & Smith, K. (2013). Communication leads to the emergence of sub-optimal category structures. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 1312–1317). Austin, TX: Cognitive Science Society.
- Silvey, C., Kirby, S., & Smith, K. (2015). Word meanings evolve to selectively preserve distinctions on salient dimensions. *Cognitive Science*, 39, 212–226. doi:10/f6x6t7
- Sims, C. R. (2018). Efficient coding explains the universal law of generalization in human perception. *Science*, 360, 652–656. doi:10/gdjs9j
- Smith, K. (2004). The evolution of vocabulary. *Journal of Theoretical Biology*, 228, 127–142. doi:10/czd9gq
- Smith, K., & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition*, 116, 444–449. doi:10/d7rxt5
- Solomonoff, R. J. (1964a). A formal theory of inductive inference. Part I. *Information and Control*, 7, 1–22. doi:10/css54w
- Solomonoff, R. J. (1964b). A formal theory of inductive inference. Part II. *Information and Control*, 7, 224–254. doi:10/d4scwz
- Solomonoff, R. J. (1997). The discovery of algorithmic probability. *Journal of Computer and System Sciences*, 55, 73–88. doi:10/b58bjk
- Spike, M., Stadler, K., Kirby, S., & Smith, K. (2017). Minimal requirements for the emergence of learned signaling. *Cognitive Science*, 41, 623–658. doi:10/f954rm
- Stadler, K. (2017). *The Page test is not a trend test* (Tech. Rep.). Retrieved from <https://kevinstadler.github.io/cultevo/articles/page.test.html>



- Steels, L. (1995). A self-organizing spatial vocabulary. *Artificial Life*, 2, 319–332. doi:10/bsjctv
- Steinert-Threlkeld, S., & Szymanik, J. (under review). Ease of learning explains semantic universals. Retrieved from <https://semanticsarchive.net/Archive/zM5ZGIxM/EaseLearning.pdf>
- Suffill, E., Branigan, H., & Pickering, M. J. (2016). When the Words Don't Matter: Arbitrary labels improve categorical alignment through the anchoring of categories. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 1715–1720). Austin, TX: Cognitive Science Society.
- Szabó, Z. G. (2013). Compositionality. In E. N. Zalta (Ed.), *The Stanford encyclopedia of Philosophy* (Fall 2013 ed.). Retrieved from <http://plato.stanford.edu/entries/compositionality/>
- Tamariz, M. (2008). Exploring systematicity between phonological and context-cooccurrence representations of the mental lexicon. *The Mental Lexicon*, 3, 259–278. doi:10/fk6bmt
- Tamariz, M. (2017). Experimental studies on the cultural evolution of language. *Annual Review of Linguistics*, 3, 389–407. doi:10/gfr9g3
- Tamariz, M., Kirby, S., & Carr, J. W. (2016). Cultural evolution across domains: Language, technology and art. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 2759–2764). Austin, TX: Cognitive Science Society.
- Thompson, B., Kirby, S., & Smith, K. (2016). Culture shapes the evolution of cognition. *Proceedings of the National Academy of Sciences of the USA*, 113, 4530–4535. doi:10/f8kzg4
- Thompson, P. D., & Estes, Z. (2011). Sound symbolic naming of novel objects is a graded function. *The Quarterly Journal of Experimental Psychology*, 64, 2392–2404. doi:10/d6vjcb
- Tinits, P., Nölle, J., & Hartmann, S. (2017). Usage context influences the evolution of overspecification in iterated learning. *Journal of Language Evolution*, 2, 148–159. doi:10/gdm73w
- Tishby, N., Pereira, F. C., & Bialek, W. (1999). The information bottleneck method. In

- The 37th Annual Allerton Conference on Communication, Control, and Computing* (pp. 368–377).
- Tversky, A. (1977). Features of similarity. *Psychological Review*, 84, 327–352. doi:10/ddf9w5
- Verhoef, T. (2012). The origins of duality of patterning in artificial whistled languages. *Language and Cognition*, 4, 357–380. doi:10/gd8c4j
- Verhoef, T., Kirby, S., & de Boer, B. (2015). Iconicity and the emergence of combinatorial structure in language. *Cognitive Science*, 40, 1969–1994. doi:10/f9cz39
- von der Gabelentz, G. (1891). *Die Sprachwissenschaft: Ihre Aufgaben, Methoden und bisherigen Ergebnisse* [Linguistics: Aims, methods, and current results]. Leipzig, Germany: T.O. Weigel Nachfolger.
- Watson, D. F. (1981). Computing the  $n$ -dimensional Delaunay tessellation with application to Voronoi polytopes. *The Computer Journal*, 24, 167–172. doi:10/bf4kcs
- Welch, T. A. (1984). A technique for high-performance data compression. *Computer*, 17, 8–19. doi:10/bf2vg5
- Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences of the USA*, 104, 7780–7785. doi:10/bms49k
- 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, 503–513. doi:10/f8w9zt
- Winter, B., & Wieling, M. (2016). How to analyze linguistic change using mixed models, Growth Curve Analysis and Generalized Additive Modeling. *Journal of Language Evolution*, 1, 7–18. doi:10/gfr9g4
- Winters, J., Kirby, S., & Smith, K. (2015). Languages adapt to their contextual niche. *Language and Cognition*, 7, 415–449. doi:10/gfr9g5
- Winters, J., Kirby, S., & Smith, K. (2018). Contextual predictability shapes signal autonomy. *Cognition*, 176, 15–30. doi:10/gdp5bp
- Wolff, J. G. (1982). Language acquisition, data compression and generalization. *Language and Communication*, 2, 57–89. doi:10/bpm8fq
- Wray, A., & Perkins, M. (2000). The functions of formulaic language: An integrated model. *Language and Communication*, 20, 1–28. doi:10/bkhp3

- Xu, J., Dowman, M., & Griffiths, T. L. (2013). Cultural transmission results in convergence towards colour term universals. *Proceedings of the Royal Society B: Biological Sciences*, 280, 1–8. doi:10/gfr9g7
- Xu, Y., & Regier, T. (2014). Numeral systems across languages support efficient communication: From approximate numerosity to recursion. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 1802–1807). Austin, TX: Cognitive Science Society.
- Xu, Y., Regier, T., & Malt, B. C. (2016). Historical semantic chaining and efficient communication: The case of container names. *Cognitive Science*, 40, 2081–2094. doi:10/f9c2j6
- Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences of the USA*, 115, 7937–7942. doi:10/gd2ssv
- Zenil, H., Soler-Toscano, F., Delahaye, J.-P., & Gauvrit, N. (2015). Two-dimensional Kolmogorov complexity and an empirical validation of the Coding theorem method by compressibility. *PeerJ Computer Science*, 1, 1–31. doi:10/gd3tcf
- Zenil, H., Soler-Toscano, F., Dingle, K., & Louis, A. A. (2014). Correlation of automorphism group size and topological properties with program-size complexity evaluations of graphs and complex networks. *Physica A: Statistical Mechanics and its Applications*, 404, 341–358. doi:10/f25fx4
- Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology*, 111, 493–504. doi:10/f854b9
- Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge, MA: Addison–Wesley.
- Ziv, J., & Lempel, A. (1978). Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, 24, 530–536. doi:10/b4c9q8