



THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Context, Cognition and Communication in Language

James Winters

Doctor of Philosophy
School of Philosophy, Psychology, and Language Sciences
University of Edinburgh
2016

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

(James Winters)

Abstract

Questions pertaining to the unique structure and organisation of language have a long history in the field of linguistics. In recent years, researchers have explored cultural evolutionary explanations, showing how language structure emerges from weak biases amplified over repeated patterns of learning and use. One outstanding issue in these frameworks is accounting for the role of context. In particular, many linguistic phenomena are said to be context-dependent; interpretation does not take place in a void, and requires enrichment from the current state of the conversation, the physical situation, and common knowledge about the world. Modelling the relationship between language structure and context is therefore crucial for developing a cultural evolutionary approach to language. One approach is to use statistical analyses to investigate large-scale, cross-cultural datasets. However, due to the inherent limitations of statistical analyses, especially with regards to the inadequacy of these methods to test hypotheses about causal relationships, I argue that experiments are better suited to address questions pertaining to language structure and context. From here, I present a series of artificial language experiments, with the central aim being to test how manipulations to context influence the structure and organisation of language. Experiment 1 builds upon previous work in iterated learning and communication games through demonstrating that the emergence of optimal communication systems is contingent on the contexts in which languages are learned and used. The results show that language systems gradually evolve to only encode information that is informative for conveying the intended meaning of the speaker – resulting in markedly different systems of communication. Whereas Experiment 1 focused on how context influences the emergence of structure, Experiments 2 and 3 investigate under what circumstances do manipulations to context result in the loss of structure. While the results are inconclusive across these two experiments, there is tentative evidence that manipulations to context can disrupt structure, but only when interacting with other factors. Lastly, Experiment 4 investigates whether the degree of *signal autonomy* (the capacity for a signal to be inter-

preted without recourse to contextual information) is shaped by manipulations to *contextual predictability*: the extent to which a speaker can estimate and exploit contextual information a hearer uses in interpreting an utterance. When the context is predictable, speakers organise languages to be less autonomous (more context-dependent) through combining linguistic signals with contextual information to reduce effort in production and minimise uncertainty in comprehension. By decreasing contextual predictability, speakers increasingly rely on strategies that promote more autonomous signals, as these signals depend less on contextual information to discriminate between possible meanings. Overall, these experiments provide proof-of-concept for investigating the relationship between language structure and context, showing that the organisational principles underpinning language are the result of competing pressures from context, cognition, and communication.

Acknowledgements

A thesis is very rarely the product of a single individual. Like language, it gradually emerges from repeated interactions – and this thesis is no exception.

Special thanks go to my two supervisors, Simon Kirby and Kenny Smith, for giving me the opportunity to do a PhD in the first place and for their support and advice throughout my time at Edinburgh. You guys are a continual source of intellectual inspiration. (Additional thanks to Kenny for continuing the tradition of explaining the difference between a shovel and a spade to this naïve learner.)

For being a long-term collaborator, Replicated Typo co-founder, groomsman, and, most importantly, a good friend, I thank Seán Roberts.

Without the financial support of the AHRC this PhD could not have happened. I hope I was a good investment.

There are many others who have contributed in various ways to this thesis over the last five years. I would like to thank, wholeheartedly:

The inhabitants of office 1.15 (both past and present): Mark Atkinson, Kevin Stadler, Matt Spike, Catriona Silvey, Yasamin Motamedi, Carmen Saldaña, Marieke Woensdregt, Thijs Lubbers, Vanessa Ferdinand, Bill Thompson, Jasmeen Kanwal, Alan Nielsen, George Starling, Steph DeMarco, James Thomas, Justin Sulik, Soundness Azzabou-Kacem, Rea Colleran, Jon Carr, Daniel Lawrence, Andrea Ravnani, Ashley Micklos, Christoph Hesse, Elyse Jamieson, Cathleen O’Grady, and John-Sebastian Schutter.

The various individuals involved with the LEC (now CLE) in some capacity or another, especially: Thom Scott-Phillips, Monica Tamariz, Graeme Trousdale, Christine Cuskley, Chris Cummins, Rob Truswell, Richard Blythe, Jennifer Culbertson, Jim Hurford, Olga Fehér, Marieke Schouwstra, Tessa Verhoef, Gareth Roberts, and Hannah Cornish.

My new academic home at the Mint: Olivier Morin, Piers Kelly, Thomas Müller, Lidiia Romanova, and Barbara Pavlek.

The readers of Replicated Typo and the contributors who make it possible (I promise that I'll spend more time writing posts now that this thesis is finished): Michael Pleyer, Stefan Hartmann, Hannah Little, Anne Pritchard, Bill Benzon, Christian Kliesch, Jess Hrebinka, Keelin Murray, Richard Littauer, and Rachael Bailes.

For unbounded support (both emotional and financial) throughout the years I would like to thank my family, especially my Mum, Dad, and my brother Chris.

Finally, for keeping me sane over the last few years, being continually patient, and making the insane commitment to marry a penniless and stressed PhD student, I would like to thank my wife, Maria Brackin. You make it all worthwhile.

Contents

Abstract	6
1 Introduction	11
1.1 Introduction	11
1.2 What is context?	13
1.3 Filling the expressive gaps	17
1.4 The interaction of minds and data	18
1.4.1 Discrimination Pressure: Context and Expressivity	22
1.4.2 Generalisation Pressure: Cognition and Simplicity	24
1.4.3 Coordination Pressure: Communication and Convention	26
1.5 Methodological Commitments	30
1.6 Thesis road map	31
2 Languages adapt to their contextual niche	33
2.1 Introduction	33
2.2 Author contributions	34
2.3 Winters, Kirby & Smith (2015): Languages adapt to their contextual niche	34
2.4 Conclusion	71
3 From <i>cup board</i> to <i>cupboard</i>: The loss of compositionality in linguistic systems	73
3.1 Introduction	73
3.1.1 Loss of compositionality: reuse, chunking, and context	75
3.1.2 Modelling the loss of compositionality	78
3.2 Experiment 2	82
3.2.1 Method	82
3.2.2 Dependent Variables and Hypotheses	84
3.2.3 Results	86
3.2.4 Experiment 2 Discussion	90

3.3	Experiment 3	95
3.3.1	Shape Bias or Unbalanced Dimensions?	95
3.3.2	Method	95
3.3.3	Results	97
3.3.4	Experiment 3 Discussion	102
3.4	Design Issues and Future Directions	106
3.4.1	Changes to the meaning space	106
3.4.2	Changes to the context	107
3.4.3	Changes to the task	107
3.5	Conclusion	108
4	Signal autonomy is shaped by contextual predictability	111
4.1	Introduction	111
4.2	Author contributions	112
4.3	Winters, Kirby & Smith (submitted): Signal autonomy is shaped by contextual predictability	112
4.4	Conclusion	170
5	Linguistic diversity and traffic accidents	171
5.1	Introduction	171
5.2	Author contributions	173
5.3	Roberts & Winters (2013): Linguistic diversity and traffic accidents	173
5.4	Conclusion	187
6	Conclusion	189
	Appendices	241
	Appendix A: Languages Experiment 1	241
	Appendix B: Languages Experiment 2	243
	Appendix C: Languages Experiment 3	244
	Appendix D: Languages Experiment 4	245

Chapter 1

Introduction

1.1 Introduction

“Why did you write that?” asked the supervisor. “Because I thought it would be a good example,” replied the student. What at first might seem like an unremarkable and commonplace interaction is in fact a demonstration of an ability found only in humans (Pinker, 1994). By using a set of sounds, strung together in a particular sequence, the student and supervisor are able to convey their thoughts to one another. This capacity to express almost any thought, and have someone else understand your expression and reply in kind, reveals three fundamental aspects of what we call *language*. First, language is extremely productive, in that it is capable of expressing novel thoughts. Second, the expressions generated are not simply random sequences; they are systematic and rule-governed. Lastly, even in the simple interaction above, to fully understand the supervisor’s question, as well as the student’s response, you need to know the context in which the expression is situated.

Questions pertaining to the unique structure and organisation of language have a long history in the field of linguistics and philosophy (von Humboldt, 1836; Frege, 1884; Paul, 1897; de Saussure, 1916; Sapir, 1921; Wittgenstein, 1921; Bloomfield, 1933; Chomsky, 1957; Hockett, 1960; Greenberg, 1963). One question which dominates many of these works is: Why is language structured in a certain way and not another? This ultimately lends itself to the evolutionary question of how language emerged in the first place and to what extent the fit between form and function is the result of biological and cultural processes (Darwin, 1871; Chomsky, 1988; Gould, 1987; Pinker & Bloom, 1990; Christiansen & Kirby, 2003). In recent years, cultural evolutionary explanations have made considerable progress in exploring the origins and evolution of language, showing how

language structure emerges from weak biases amplified over repeated patterns of learning and use (for a recent review, see: Tamariz & Kirby, 2016). However, little attention has been paid to the role of context in the cultural evolution of language – and how it interacts with cognition and communication (see, e.g., Steels, 1999; Silvey, 2015 for notable exceptions).

Yet, knowledge of context is clearly a crucial component in how humans produce and comprehend language, as identified by work in psycholinguistics (for review, see: Konopka & Brown-Schmidt, 2014), pragmatics (for review, see: Attardo, 2016), historical linguistics (for review, see: Traugott & Trousdale, 2013), and cognitive linguistics (for review, see: Evans & Green, 2006). The central aim of this thesis is to investigate the causal relationship between context, cognition, and language structure using a cultural evolutionary framework. What constitutes a context, and how it interacts with the structure and use of language, is still subject to considerable debate (Terkourafi, 2009; Faber & León-Araúz, 2016). As Herb Clark noted in motivating the use of his term *common ground*:

In the study of language use, investigators appeal time and again to “context” to explain this or that phenomenon. The problem is that they almost never say what they mean by “context” even when it is essential to their explanations [...] it also allows psychologists to hoodwink themselves as well as others into thinking they have explanations for context effects when they don’t. I prefer to eschew the term *context* for something less dangerous. (Clark, 1992: 6; emphases in original).

Taking Clark’s point seriously, but opting for what is perhaps the more dangerous path, I am going to stick with the term context and argue it can be a useful explanatory device when approached correctly. The first step, then, is to come up with a precise definition of context which can be operationalised, separated from other causal factors, and investigated in a systematic fashion. In particular, I will advocate the use of laboratory experiments, drawing on a growing body of work in iterated learning and communication games. Experiments represent a natural middle ground between the abstractness and manipulability of computational or mathematical simulations and the need for observable, real world data found in corpora. Through using experiments I will investigate how context interacts with learning and communication to shape the evolution of distinct communication systems; how context constrains the learning and use of compositional mappings; and how manipulations to the context influence a speaker’s ability to estimate, and therefore exploit, the information that a hearer is likely to use in interpreting

an utterance. Parcelling out the effects of context from other constraints found in learning and communication allows us to investigate how these factors interact with one another in shaping the structure of language.

The subsequent sections of this introduction will ask what context is, how it gets into the structure of language, and motivate the use of experiments as a means of investigating the link between context and language structure. Lastly, I will provide a roadmap for the rest of the thesis.

1.2 What is context?

As a concept, *context* is found in numerous disciplines, ranging from psychology (Todorović, 2010) and linguistics (Faber & Leon-Arauz, 2016) to architecture (Hinton, 2014) and computer science (Bazire & Brezillon, 2005). So, when presented with the question *what is context?*, it should not come as a surprise that a clear cut answer is hard to come by – definitions are often subject to book-length treatments (e.g., Givon, 2005; Bergs & Diewald, 2009; Finkbeiner, Meibauer & Schumacher, 2012) and vary both between and within disciplines (Bazire & Brezillon, 2005). Providing a consensual definition often results in a concept so vague and nebulous that “context can refer to the whole universe” (Fetzer, 2004: 3), is a term “which is constantly used in all kinds of context but never explained” (Quasthoff, 1998: 157) and acts as “conceptual garbage can” (Smith, Glenberg & Bjork, 1978: 342). This is pretty much what we get when, based on a corpus of 150 definitions, the notion of context is distilled down into several general parameters: “[context is] a set of *constraints* that *influence* the *behavior* of a *system* (*a user or a computer*) *embedded in a given task*” (Bazire & Brezillon, 2005: 39; emphases in original).

As the authors of the above quote concede, this definition of context fails to address some of the key debates surrounding the term, and does not tell us what these constraints are and how they influence the behaviour of the system. For investigating the relationship between context and language structure this is not a very useful definition. An alternative starting point comes from Sperber & Wilson:

A context is a *psychological construct*, a subset of the hearer’s assumptions about the world. It is these assumptions, of course, rather than the actual state of the world, that affect the interpretation of an utterance. A context in this sense is not limited to information about the immediate physical environment or the immediately preceding

utterances: expectations about the future, scientific hypotheses or religious beliefs, anecdotal memories, general cultural assumptions, beliefs about the mental state of the speaker, may all play a role in interpretation. (Sperber & Wilson, 1986: 15; my emphasis).

Like other definitions, Sperber & Wilson include an expansive list of properties which may be considered part of the context, but the important take-home message is their approach to context as a *psychological construct* (for review, see: Attardo, 2016). Similar approaches are found in Baars (1998), who thinks of the context as a mental phenomenon more accurately located inside the nervous system, and Clark & Carlson’s (1992: 65) view of context as the “information that is available to a particular person for interaction with a particular process on a particular occasion”. Thinking of context in this way disentangles the definition from being tied to an external property of the world, synonymous with the environment, and firmly places it in the minds of individuals. And, as a psychological construct, context is very much embedded in the business of cognition: to generate predictive models of the world (Clark, 2015).

To illustrate the relationship between context, cognition, and prediction consider the 13/B visual illusion in Figure 1.1. When reading from top to bottom, the central object is interpreted as the number 13, whereas reading from left to right results in interpreting the central object as the letter B. This simple illusion demonstrates that placing the same object in different situations can have a profound impact on our interpretation. Viewed in this light, context links together cognition, perception, and the environment in generating predictions via a *frame of interpretation* (Goffman, 1974; Minsky, 1975; Fauconnier, 1985; Fillmore, 1985): defined in this thesis as *knowledge derived from perceivable relationships in the environment used in generating a prediction*.



Figure 1.1: The 13/B illusion. The left image highlights the *number context* and the right image highlights the *alphabet context*.

Treating context as a frame of interpretation allows us to break it down into three key components: the figure (target of interpretation), the ground (the immediate information brought to bear to the task of interpretation), and the background (prior knowledge which helps delineate what enters into a frame) (Duranti & Goodwin, 1992; Terkourafi, 2009). In the case of the 13/B illusion, the figure is the central object, the ground is either the *letter context* or *alphabet context*, and the background is prior knowledge about numbers and letters. Arriving at an interpretation is therefore contingent on the contextual frame and the way it integrates what is currently predicted with incoming sensory data (Clark, 2015). As such, generating a predictive model depends not only on the environment in which the object is situated – it is also the cognitive and perceptual apparatus of the individual engaged in a task and how these factors conspire together to create a frame.

The first point to note is that the contextual frame directs attention: it highlights and backgrounds information in making a prediction. In order to do this, an organism must be capable of detecting and extracting information from latent structures in the environment, through “narrowing down from a vast manifold of information to the minimal, optimal information that specifies the affordance of an event, object, or layout” (Gibson & Pick, 2000: 149). For the 13/B illusion, an individual must use their perceptual apparatus to perceive character strokes, and integrate these strokes to form a holistic character, before they can differentiate and learn the relationships between a sequence of these characters.

This leads us to our second point: what enters into a frame also consists of prior knowledge about previous frames. This creates a causal link where previous frames, and the interpretations derived from those frames, influence the creation of subsequent frames. Knowledge here refers to stored tokens of experience which are categorised and matched with similar tokens to form exemplars (Evans & Green, 2006; Elman, 2009; Bybee, 2010; Port & Ramscar, 2015). For example, based on our experience of the English alphabet, we store information about how the letter *B*, when arranged in a certain sequence (alphabet context), normally follows the letter *A* and precedes the letter *C*. A contextual frame is therefore built up over repeated experiences, refined through feedback from prior predictions, and then used in generating the current prediction: if a literate English speaker sees a character situated between the letters *A* and *C* there is a high probability this is the letter *B*. And it is this probability which creates a strong bias in making a prediction.

All this leads to the general conclusion where establishing a contextual frame creates a *discrimination pressure*: determining what is and is not informative

in reducing uncertainty in interpretation. Organisms fine-tune their predictive models of the world through integrating information from the contextual frame to discriminate against uninformative cues (those that do not improve predictions) and reinforce informative cues (those which tend to improve predictions) (Ramscar & Port, 2015). By deliberately attending to some information, and learning to ignore information irrelevant to the task, an organism is able to improve its predictions over time. Such strategies are widely-acknowledged in the decision-making literature where situations characterised by a high degree of uncertainty are more accurately solved using less information or computation (e.g., *less-is-more effects*: Gigerenzer & Brighton, 2009; Gigerenzer & Gaissmaier, 2011).

To spell out how contextual information becomes embedded in a system of behaviour imagine a rat placed in a T-Maze: the contextual frame consists of a set of options (in this case, the left arm or the right arm), the rat's ability to perceive the relationship between these options, and its prior knowledge about which of these options contains a reward. Making a decision about which arm contains a reward is at chance in the first trial: the rat simply knows there are two options and it makes a guess as to the location of the reward. Even in this simple situation the rat is attending to salient aspects of the environment (e.g., the two options) and ignoring irrelevant information (e.g., the colour of the maze walls). Over successive trials, the rat learns that the reward is consistently in the left arm, and adjusts its behaviour to always go to this location. If the reward is now switched to the right arm for several trials, then the rat uses this new data to update its predictions, and gradually shifts from a left-arm to a right-arm preference. By continually switching the reward arm the rat becomes more efficient in updating its predictions and incorporates the contextual cues into its model: if the rat predicts the reward is in the left arm, and this prediction turns out to be incorrect, then in the next trial the rat predicts the reward is in the right arm (*serial reversal-learning paradigm*; for review, see: Lloyd & Leslie, 2013).

As the rat example illustrates, discrimination is inherent to any perceptual task (Wade & Swanston, 2001), and is thus independent from language and communication¹. By providing a clearer definition of context (as a frame of interpretation), and how it relates to cognition and behaviour, we can now move onto the issue of the relationship between context and the structure of language.

¹But as we shall see the pressure to discriminate can be motivated by communicative needs. The point here is to separate the capacity for discrimination, which is broadly a perceptual propensity found in numerous cognitive domains, from our capacity to use language as a tool for communication.

1.3 Filling the expressive gaps

Like other behaviours, linguistic structure appears to rely on contextual information for disambiguation. Consider how even a simple English sentence, such as *she passed the mole*, is ambiguous as to whether the verb *passed* refers to *a form of motion* or *an act of giving*, and whether the noun *mole* refers to a *small burrowing mammal*, *a person engaged in espionage*, *a brand of Mexican sauce* or *a type of causeway*. Context in this sense could be as immediate as the surrounding linguistic information (e.g., *Maria was riding her bike and she passed the mole which was was burying into the ground*) and as remote as knowledge about the prior discourse (e.g., the story is about cold war spies).

What both of these aspects share is that they form a frame of interpretation. This is important because it focuses on the effects of context – to create a discrimination pressure which determines what is and is not informative for interpretation – and less on listing all of the possible types of context. In short, contextual information increases the probability of some interpretations over others (for more details on referential context and ambiguity resolution, see: Haywood, Pickering & Branigan, 2005; Spivey-Knowlton & Tanenhaus, 2015). Still, we need to at least establish what the relationship is between language use, language structure and context, before we can move on to discuss how and why this relationship emerged in the first place. Indeed, the fact that contextual information is needed to resolve communicative issues, such as the presence of ambiguity, is often held up as an example of how language is poorly designed with respect to its communicative utility (e.g., Berwick et al., 2011):

The natural approach has always been: is it [language] well designed for use, understood typically as use for communication? I think that's the wrong question. The use for communication might turn out to be a kind of epiphenomenon. I mean, the system developed however it did, we really don't know. And then we can ask: how do people use it? It might turn out that it is not optimal for some of the ways in which we want to use it. If you want to make sure that we never misunderstand one another, for that purpose language is not well designed, because you have such properties as ambiguity. (Chomsky, 2002: 107).

Many authors have argued that Chomsky's approach is completely backwards, with ambiguity not only being an easily resolvable problem, but also an expected property of an efficient communication system which combines our powerful inferential capacities with contextual information (Piantadosi et al., 2012; Scott-Phillips, 2015). In fact, the extent to which ambiguity genuinely impedes

communication is so rare, or so quickly resolved, that an optimally efficient communication system *should* be ambiguous (Pinker & Bloom, 1990; Juba et al., 2011):

...when context is informative, any good communication will *leave out information already in the context*. . . [A]s long as there are some ambiguities that context can resolve, efficient communication systems will use ambiguity to make communication easier. (Piantadosi, Tily & Gibson, 2012: 284; emphasis is mine).

This relationship between context, language use, and language structure can be thought of as filling in expressive gaps: if the context already contains information about the intended meaning, then there is no need to explicitly express this information linguistically. This offers a functional explanation for the fit between context and language structure: language is context-dependent because it is efficient for the purposes of communication. Hermann Paul made a similar observation 126 years ago when he wrote:

The more economical or more abundant use of linguistic means of expressing a thought is determined by the need... Everywhere we find modes of expression forced into existence which contain only just so much as is requisite to their being understood. The amount of linguistic material employed varies in each case with the situation, with the previous conversation, with the relative approximation of the speakers to a common state of mind. (Paul, 1890: 251).

Still, the fact that context-dependency makes good design sense does not necessarily explain why this relationship emerged in the first place. The next section will sketch out a model where context becomes instantiated in the structure of language through the process of cultural transmission.

1.4 The interaction of minds and data

Understanding the causal relationship between context and language structure requires that we take seriously the *problem of linkage* (Kirby, 1999): How are individual behaviours linked to the way language is organised and structured? That is, given the observation of context-dependency in language, and the corresponding claim that this is communicatively efficient, we still do not know how these advantages “become realized in language as a system of human behaviour”

(Smith & Kirby, 2012: 494). An increasingly prominent approach to addressing the problem of linkage is to think of language as a *Complex Adaptive System* (henceforth, CAS) (Steels, 2000; Beckner et al., 2009):

(1) The system consists of multiple agents (the speakers in the speech community) interacting with one another. (2) The system is adaptive, that is, speakers' behavior is based on their past interactions, and current and past interactions together feed forward into future behavior. (3) A speaker's behavior is the consequence of competing factors ranging from perceptual mechanics to social motivations. (4) The structures of language emerge from interrelated patterns of experience, social interaction and cognitive processes. (Beckner et al., 2009: 3).

Crucially, language is viewed as “the interaction of minds and data” (Hurford, 2003: 51), and exists at two interdependent junctures, consisting of an *idiolect* (the set of form-meaning mappings belonging to an individual language user) and the *communal language* (the set of conventions shared by a community of language users). Both of these aspects are emergent: idiolects emerge from each individual's use of their language through repeated interactions in the speech community, and the communal language is a product of negotiated interactions between these idiolects (Beckner et al., 2009). This feedback loop between idiolects and the communal language means that the link between language structure and individual behaviours is characterised by an additional dynamic system: *cultural transmission* (Kirby & Hurford, 2002; Kirby et al., 2007; Christiansen & Chater, 2008).

Cultural transmission forms the backbone of theories where language is the result of cultural evolutionary processes (Smith & Kirby, 2012). As with biological systems, language and other cultural systems are considered evolutionary because they meet the requirements of reproduction, heritable variation, and differential amplification of variants (Boyd & Richerson, 1985²; Croft, 2000; Mufwene, 2001; Ritt, 2004; Steels, 2012; Kirby, 2013). Unlike biological evolution, where inheritance involves the direct replication of DNA, the cultural transmission of language is indirect and takes place through a process known as *iterated learning*: “by which a behaviour arises in one individual through induction on the basis of observations of behaviour in another individual *who acquired that behaviour in the same way*” (Kirby, Griffiths, & Smith, 2014: 108; emphasis in original).

²Boyd & Richerson (1985) provide a general approach to cultural evolution that employs methods from population genetics to investigate the effects of psychological biases on cultural variants. Also see Mesoudi (2011) for a relatively up-to-date overview of research into cultural evolution.

Treating language as a culturally transmitted system solves the problem of linkage through showing how short-term behaviours, used in solving immediate communicative needs, are amplified over repeated interactions to shape sets of behaviours shared by a population of individuals. What are the individual behaviours in language? What is transmitted during these interactions? The key unit of individual behaviour which is transmitted from speakers to hearers³ is the *utterance*: “a situated instance of language use which is culturally and contextually embedded and represents an instance of linguistic behaviour on the part of a *language user*”. (Evans & Green, 2006: 110; emphasis in original).

As the above quote indicates, an utterance is strongly coupled to its contexts of use, and recognises that language behaviour is fundamentally communicative. This introduces an important distinction in discussing context and its relationship to language when compared to context and other behaviours involving perception:

The intrinsic context [the directly relevant information for completing a specific task] for comprehension is different in one fundamental way from most other notions of intrinsic context. In areas like visual perception, the notion of common ground isn’t even definable, because there are generally no agents involved other than the perceiver himself. Defining the intrinsic context in terms of common ground appears to be limited to certain processes of communication. Context, therefore, cannot be given a uniform treatment across all psychological domains. In language comprehension, indeed, the intrinsic context is something very special. (Clark & Carlson, 1992: 77).

So, whereas the role of context in many perceptual domains is to coordinate the behaviour of individuals with the world, language is special in that individuals are also trying to coordinate with one another (Chater & Christiansen, 2010). That is, in producing and comprehending an utterance, speakers and hearers engage in a collaborative activity where they draw on shared situational and world knowledge in order to align on similar frames of interpretation (Clark, 1992; Parikh, 2001; Franke, 2013). If we consider meaning to simply reside in the act of interpretation, and context is the information which contributes to this interpretation by creating a frame, then this provides a mechanism through which utterances inherit interpretations as they are produced and perceived during transmission.

³A hearer can be anyone who is capable of learning and using a language. This allows us to reconcile any differences which exist between transmission between already-competent users of language and naïve learners (for a discussion on these differences, see: Labov, 2007; Lupyan & Dale, 2010; Kirby et al., 2015).

Under this account, utterance meaning is not really “prepackaged chunks of information” (Geeraerts, 1993: 263), with linguistic coding being “less like definitive content and more like interpretive clue” (Levinson, 2000: 29). Importantly, the production and comprehension of these utterances is governed by competing constraints, ranging from the organisation of thought processes and perceptuo-motor machinery to cognitive and pragmatic factors (Christiansen & Chater, 2008; Chater & Christiansen, 2010). In this sense, a constraint is essentially a canalising factor – limiting the search space in which behavioural variation is allowed to explore. What enters into a contextual frame during language use is therefore contingent on the constraints inherent to the transmission of utterances.

Cultural evolutionary approaches have been particularly successful in demonstrating how competing constraints interact over multiple timescales (Kirby et al., 2007; Smith, 2009; Kirby et al., 2015; Thompson, Kirby & Smith, 2016). Much of the focus in these works has been on two constraints: the first is a domain-independent cognitive bias for *simplicity* (Chater & Vitanyi, 2003; Kemp & Regier, 2012; Clark, 2015; Culbertson & Kirby, 2016) and the second is a task-specific bias of *expressivity* (Pinker & Bloom, 1990; Kirby & Hurford, 2002; Fay, Garrod & Roberts, 2008; Frank & Goodman, 2012; Piantadosi et al., 2012; Kemp & Regier, 2012; Kirby et al., 2015)⁴. A bias for simplicity reduces the algorithmic complexity of a system by making it more compressible (Tamariz & Kirby, 2014), i.e., the description length of the system is shorter than a list of the possible signal-meaning mappings. An expressivity bias corresponds to the task-specific goal of communication: to reduce uncertainty about the intended meaning in context. Language structure emerges from the interplay between these two biases as utterances are transmitted from speakers to hearers, with the effects of these biases being instantiated as system-wide characteristics:

language learning by naïve individuals introduces a pressure for simplicity arising from a domain-independent bias for compressibility in learning, and a pressure for expressivity arises from language use in communication. Crucially, both must be in play: neither pressure alone leads reliably to structure. The structural design features of language are a solution to the problem of being compressible and

⁴Many of these authors do not use the term *expressivity*, but use a closely related term such as *informativeness* (Kemp & Regier, 2012; Frank & Goodman, 2014), *accuracy* (Fay, Garrod & Roberts, 2008), or *clarity* (Piantadosi et al., 2012). Furthermore, the definition of *expressivity* has subtly shifted from earlier works, where it originally corresponded to a motivation to convey all possible distinctions in a meaning space (Kirby & Hurford, 2002; Kirby et al., 2008), to its current incarnation: to discriminate an intended referent from possible alternative referents in a context (Kirby et al., 2015).

expressive, a solution delivered by the process of cultural evolution. (Kirby et al., 2015: 88).

Languages shaped solely by the pressures found in learning result in *degenerate* structure where every possible meaning is conveyed using a single, maximally ambiguous signal (see Kirby et al., 2008 for an experimental demonstration). Conversely, when using languages to communicate, the bias for expressivity is amplified and leads to the maintenance of *holistic* languages (i.e., an incompressible system where its description length is equal to simply listing all of the form-meaning mappings). The tradeoff between these two pressures results in the emergence of *structured* languages: a compressible set of form-meaning mappings which are functional for the task of communication.

The central goal of this thesis is to investigate how context influences the tradeoff between these pressures found in learning and communication. For our purposes, we can ask three questions relating context to language: (i) What distinctions does a system need to make in order to identify the intended meaning in context (*discrimination pressure*)? (ii) How generalisable is the system to new meanings and contexts (*generalisation pressure*)? (iii) How aligned is the speaker’s intended meaning with the hearer’s interpretation (*coordination pressure*)?

1.4.1 Discrimination Pressure: Context and Expressivity

The picture presented so far places context as a frame of interpretation that governs the *discrimination pressure*: determining what is and is not informative for reducing uncertainty in interpretation. To illustrate how this discrimination pressure interacts with language consider the simple toy world in figure 1.2: here, an individual is exposed to a set of objects (coloured shapes), with the context being knowledge about the perceivable relationship between the objects, and the task is to produce utterances which discriminates one object from another. In this case, the figure is an utterance (e.g., *miko*), the ground is the three objects in the world, and the background is prior knowledge about the utterance and how it relates to discriminating between these three objects.

There are two problems facing speakers and hearers when generating a set of utterances. First, a speaker needs to discriminate their intended referent from possible alternative referents, i.e., the set of utterances must be *expressive* (Pinker & Bloom, 1990; Kirby et al., 2015). Second, in inferring the relationship between an utterance and its intended meaning, hearers are faced with the problem of

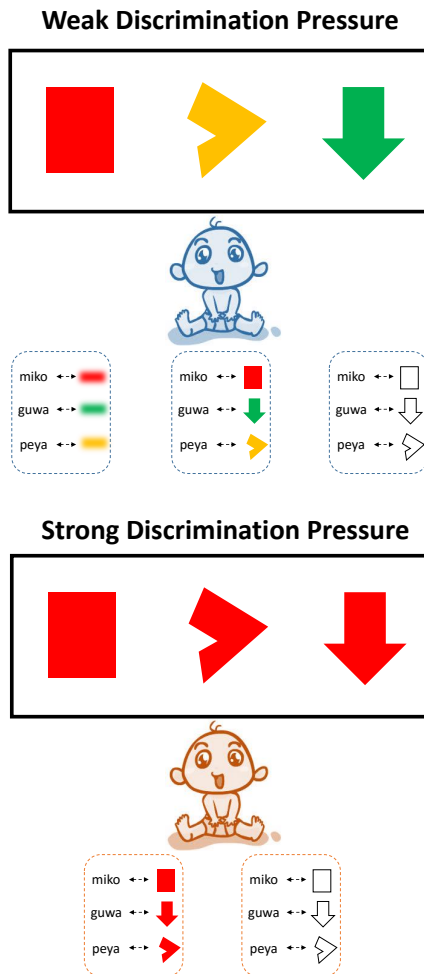


Figure 1.2: An example of a weak discrimination pressure (Blue child), where there are multiple informative dimensions, in this case colour and shape, and an example of a strong discrimination pressure (Orange child), where one dimension is informative for discrimination (shape) and the other dimension is backgrounded (colour). The task is to generate an internal system which discriminates between each of the referents in a context (the black box with three coloured shapes). Boxes with dotted lines are some of the possible sets of form-meaning mappings which can discriminate between objects given the context.

referential uncertainty (Quine, 1960; Blythe, Smith & Smith, 2016): What aspects of the referent is the utterance referring to? Context helps solve these two problems through providing information about which features are informative and uninformative for the task of discrimination. For a speaker, this translates into: maximise the capacity of an utterance to successfully convey each of the objects (Frank & Goodman, 2012). Meanwhile, hearers want utterances which, when coupled with information provided by the context, minimise the uncertainty about the intended referent (Piantadosi et al., 2012; Frank & Goodman, 2014).

Importantly, the strength of the discrimination pressure is contingent on the context: a strong discrimination pressure is created when the contextual frame unambiguously highlights which feature(s) are informative and backgrounds any which are uninformative, whereas a weak discrimination pressure arises from there being multiple informative features in the context (none of which are backgrounded). In short, information provided by the context helps rule out possible utterances, i.e., when the discrimination pressure is strong, the number of expressive utterances is lower than when the discrimination pressure is weak.

1.4.2 Generalisation Pressure: Cognition and Simplicity

Discrimination is not just a one-shot affair: speakers and hearers must produce and comprehend utterances across multiple exposures. To remain expressive a set of utterances need to resolve future predictive uncertainty (Port & Ramscar, 2015): the discrepancy between what is currently predicted and incoming sensory data (Lupyan & Clark, 2015; Clark, 2015). The challenge now is to draw on knowledge of previous interactions and use utterances which solve the immediate task at hand whilst anticipating future problems. Possible future problems in this instance include any new data (e.g., new referents, new situations etc) which decreases the probability of making successful predictions (i.e., discriminating between referents). If utterances are used only on the basis of maximising expressivity, then there is the potential for overfitting, where the set of form-meaning mappings are not generalisable to new data.

One solution to this *generalisation pressure* is to generate compressible sets of utterances. Figure 1.3 builds on the previous example to illustrate this point. Assume a speaker generates a language at $t1$ which discriminates holistically (e.g., *miko* refers only to red square). Now, for the speaker facing a weak generalisation pressure this is not an issue: the data at $t1$ are identical to the data at $t2$ and the language is adequate for the task of discrimination. However, when faced with a strong generalisation pressure, the same language is poorly equipped for discrimination, with the set of utterances generated at $t1$ not being generalisable to the new data at $t2$. As such, the strength of the generalisation pressure is tied to exposure to the data: a weak generalisation pressure is one where all future data are highly predictable, whereas a strong generalisation pressure is one where future data is associated with a high level of uncertainty. Importantly, the generalisation pressure can either be directly tied to exposure, where individuals are only exposed to a subset of the data, or indirectly related (e.g., due to general cognitive limitations on our memory and processing capabilities; see Cornish,

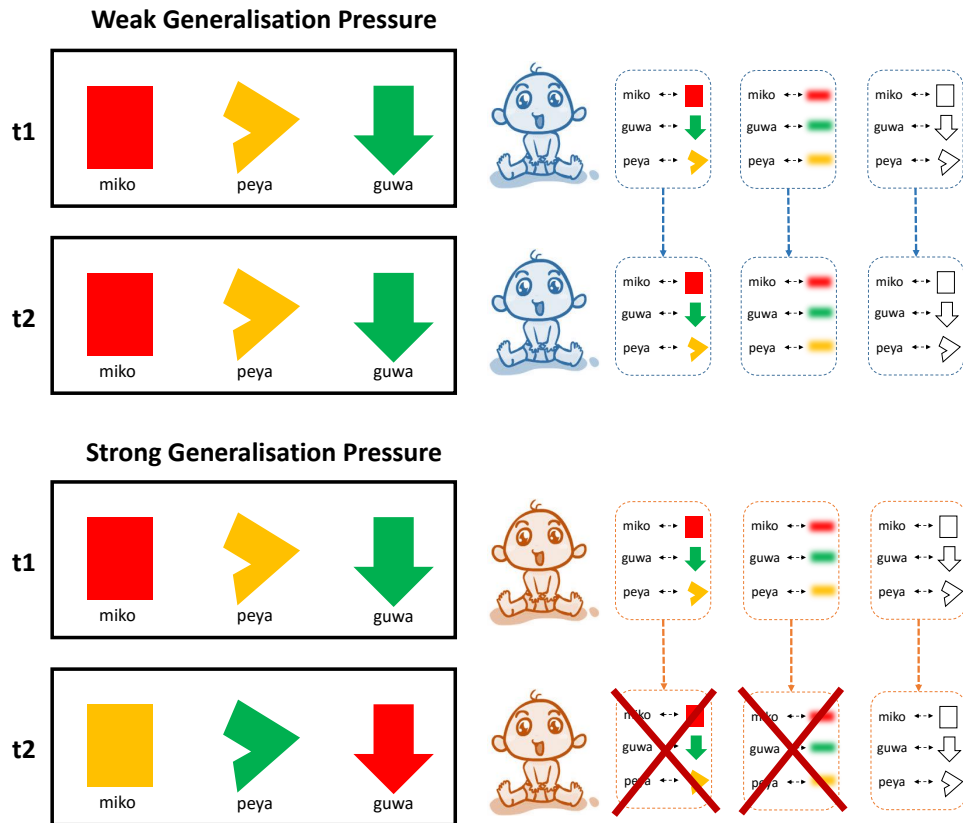


Figure 1.3: An example of a weak and strong generalisation pressure. At $t1$ and $t2$ an individual is exposed to a set of data (consisting of forms, meanings, and contexts) and must generate an internal system of form-meaning mappings. The task is to generate mappings at $t1$ which accurately discriminate between referents at $t2$ (arrows with dotted lines). In the case of the weak generalisation pressure, Blue is exposed to the same referents at $t1$ and $t2$, with any of the three systems of form-meaning mappings generated at $t1$ being adequate for discrimination at $t2$. Conversely, when faced with a strong generalisation pressure, Orange is exposed to different sets of referents at $t1$ and $t2$. The optimal approach is to use the form-meaning mappings which discriminate using shape information – allowing Orange to reuse the form-meaning mappings generated at $t1$ for $t2$. Another approach is to generate a new set of form-meaning mappings every time an individual is exposed to new data. For instance, Orange could maintain the general hypothesis that colour information is important for discrimination, but this would result in low predictive power as it entails ignoring the data at $t1$. It is also important to note that, even though the strength of discrimination pressure has not changed from $t1$ to $t2$, the strong generalisation pressure does narrow down the set of optimal form-meaning mappings.

2010; Smith & Kirby, 2012).

1.4.3 Coordination Pressure: Communication and Convention

So far, we have considered how context generates a frame of interpretation, which regulates the discrimination pressure, and how repeated exposures to these frames governs the generalisation pressure. Yet, as previously mentioned, language is not simply an internal system used by individuals (*idiolects*) – communication and social interaction are key components which cannot be ignored. If language is used for communication, then speakers and hearers need to align on shared system of communication. Without a shared system, the task of aligning on similar frames of interpretation is made extremely difficult, resulting in speakers and hearers being faced with a *coordination problem* (Lewis, 1969; Schelling, 1980; Bratman, 1992; Clark, 1996).

Coordination problems are instances where two or more individuals must solve a collaborative task by aligning on the same or corresponding strategy (Lewis, 1969). Many forms of cultural activities can be classed as coordination problems – from two dancers engaged in a tango to performers in a string quartet – but communication poses a particularly salient coordination problem: there are numerous ways in which an utterance may be interpreted and interlocutors cannot read each other’s minds (Croft, 2000). As such, every communicative situation is a recurrent coordination problem; speakers try to convey their intended meaning using an utterance, and hearers try to arrive at the correct interpretation of this utterance.

Solving the coordination problem requires speakers and hearers align on a set of conventional⁵ utterances (Lewis, 1969; Clark, 1996; Beckner et al., 2009). Establishing shared conventions is arbitrary to a certain extent, with the goal being for interlocutors to conform to what was previously negotiated, but these solutions also interact with the other pressures of discrimination and generalisation: speakers and hearers need to align on a shared system which is both expressive and compressible. The fact that these systems are negotiated between speakers and hearers in piecemeal fashion also hints at an important role for historical contingency (Lass, 1997; Millikan, 1998): previous interactions, and the solutions stemming from these interactions, constrain future outcomes. As a result, past communicative solutions can result in systems ending up in suboptimal states, where there is inertia from the set of previously established conventions:

⁵Beckner et al (2009: 4) define a convention as “a regularity of behavior (producing an utterance of a particular linguistic form) that is partly arbitrary and entrenched in the speech community”

That a token of any form instances a convention or piece of a convention is a matter of its individual history, not a matter of what it matches... Only tokens reproduced due to weight of precedent are conventional, and which convention each instances depends on the precedent from which it is derived. (Millikan, 1998: 175)

Aligning on shared communicative conventions draws upon two general resources. The first of these is feedback (De Ruiter et al., 2010; for overview, see: Spike et al., 2016): interlocutors are able to provide information to one another as to whether or not their interpretation is correct, allowing one, or both, of the parties to modify their behaviour in future interactions⁶. With feedback, speakers can modify their behaviour through trying to clarify the intended meaning, and use utterances which pick out relatively smaller sections of the context (Frank & Goodman, 2014: 85). Conversely, hearers can adjust their form-meaning mappings, aligning their interpretation with that of the speaker's intended meaning (Selten & Warglien, 2007; Moreno & Baggio, 2015).

The second of these resources is shared knowledge and the beliefs held by interlocutors about their conversational partner (Bell, 1984; Brennan & Clark, 1996; Branigan et al., 2011; Bergmann et al., 2015). Accurately assessing the degree of shared knowledge is crucial as it allows speakers to use utterances which are tailored to the needs of the hearer (Clark & Marshall, 1981⁷). For instance, when choosing referring expressions to describe a particular object, a speaker is more likely to repeat an expressions previously used by their conversational partner when they believe that partner is a computer (as humans reason they are less likely to share common ground with a computer; see Branigan et al., 2011). Similarly, if a speaker believes a hearer is part of the same community, then both individuals can assume they share some conventional knowledge, with the speaker's utterances being constructed to not only express their intended meaning, but to also indicate common ground with the hearer (e.g., contrast the use of *Empire State Building*, which relies on shared knowledge about the name of a particular building in New York City, with *the 102-story skyscraper located on Fifth Avenue between West 33rd and 34th streets*; Isaacs & Clark, 1987).

To explicitly spell out the coordination pressure, and how it interacts with the pressures of discrimination and generalisation, consider the situation in figure 1.4.

⁶Feedback is used in a very general sense here to refer any information transmitted *after* signalling that helps identify the intended meaning of the speaker and/or conveys the hearer's interpretation (Spike et al., 2016: 15).

⁷Clark & Marshall highlight three distinct sources of shared knowledge: the speech community of the interlocutors, the physical environment in which interlocutors are situated, and the internal linguistic context (e.g., the choice of words as well as the syntactic and discourse structure).

Two individuals gradually build up a system of communication and generate their own internal set of form-meaning mappings. From $t1$ to $t3$ communication proceeds unimpeded in that the set of forms are capable of discrimination and there is an underlying rule which allows for generalisation to novel objects. In these series of interactions, the coordination problem is solved through integrating contextual information into the linguistic system, allowing both Blue and Orange to ignore colour when considering which meaning they intended to convey. However, at $t4$, colour is now also informative for discrimination: to solve this coordination problem Orange draws on knowledge of previous interactions and uses a new signal to convey their intended meaning. In using a new signal, Orange is relying on the prior knowledge shared with Blue – that two of the possible referents are already associated with a conventional form – and reasons that the novel form is more likely to be associated with a novel object (in this case, a yellow square). The consequences of this interaction are made apparent at $t5$: here, Blue opts for the specific form for yellow square, as opposed to the general term for square, due to the precedent set at $t4$.

Taken together, a key issue facing this thesis is disentangling these three pressures, and how they interact with context in shaping the organisation and structure of language at multiple timescales of learning and use. The next section will outline a methodological framework for investigating the effects of context. In particular, an argument will be made for using laboratory experiments, contrasting this approach with computational modelling and large-scale corpora.

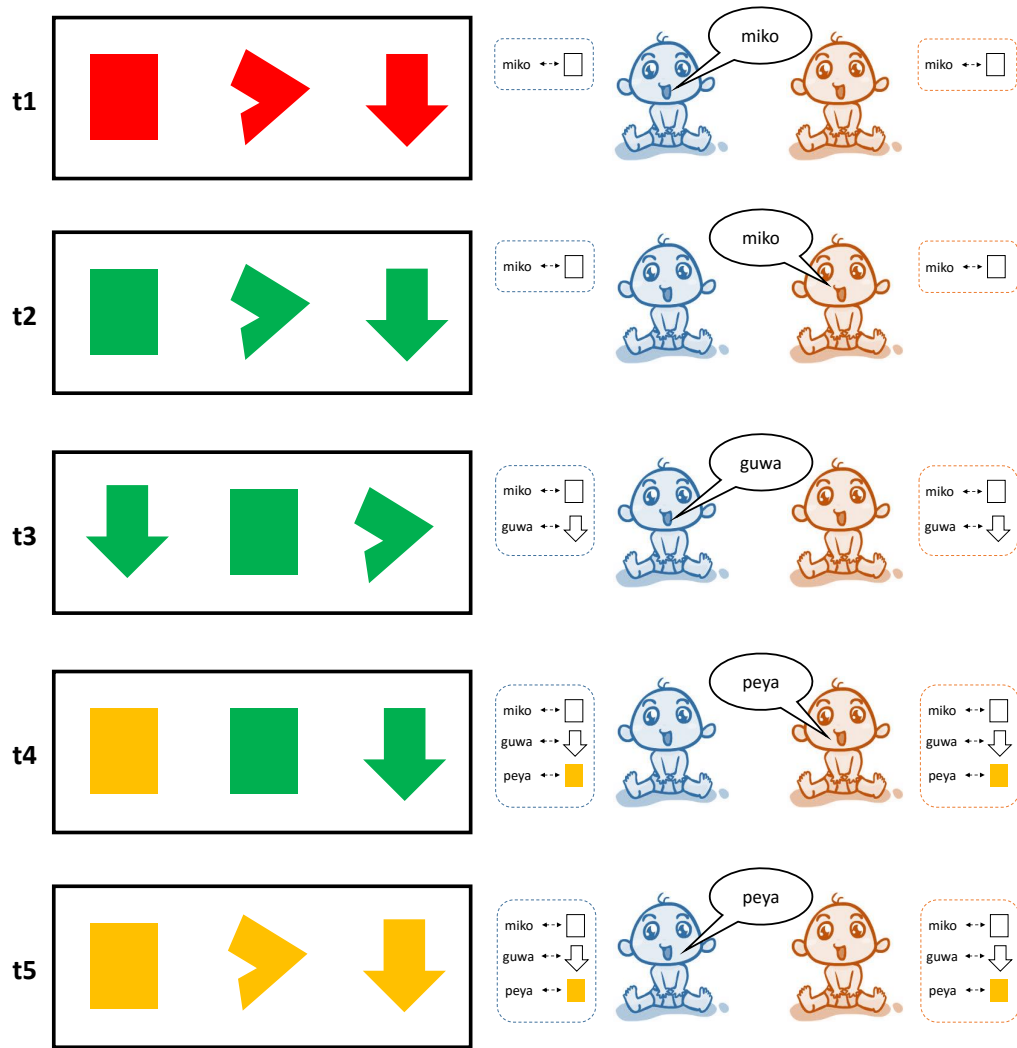


Figure 1.4: An example of how the coordination pressure might shape a communication system across a series of interactions. The task is for both individuals to successfully communicate with one another by taking turns to convey their intended meaning (which is always the object on the left). For the purposes of simplicity, the listener always correctly infers the intended meaning, even when this is just a good guess (as is the case in $t1$ where miko could have referred to any of the three objects). From $t1$ to $t3$ Blue and Orange build up a shared system of communication whereby *miko* maps onto square and *guwa* maps onto arrow (the contextual frame backgrounds colour, making it uninformative in these initial interactions). However, at $t4$, both shape and colour are informative for discrimination and Orange uses a novel signal, *peya*, to convey the yellow square. Blue is able to identify the intended meaning as they reason that the yellow square is the only novel object and Orange would have either used *miko* to refer to the green square or *guwa* to refer to the green arrow (as these are previously established mappings). The consequences of this interaction are made apparent at $t5$: here, Blue opts to use *peya*, rather than the more general *miko*, due to precedent set during the interaction at $t4$.

1.5 Methodological Commitments

There are three ways one might go about investigating the effects of context on language structure.

The first is to use statistical techniques to search for correlations in corpora and cross-linguistic databases (e.g., WALSH: Dryer & Haspelmath, 2013). Such techniques have gained in prominence over the last decade and are increasingly viewed as a viable way of investigating questions about language structure (for review, see: Ladd, Roberts & Dediu, 2015). An obvious advantage of this statistics-driven approach is that the data more closely reflects language in its natural environment – allowing us to quantify the degree of variation and similarity in how language is structured cross-linguistically. A good example of this cross-linguistic approach is Lupyan & Dale’s (2010) study into the relationship between social context and morphological complexity. Here, the authors compare the structural properties (e.g., inflectional synthesis of the verb, coding of evidentiality, number of cases etc) of more than 2000 languages with three demographic variables which act as a proxy for social context: *speaker population*, *geographic spread*, and *the number of linguistic neighbours*. In this case, certain demographic characteristics, such as a small versus a large population, are arguably useful for capturing competing pressures acting upon the structure of language. The key idea being that differences in social context correspond to variability in a society’s reliance on shared cultural and contextual information for communicating with one another (Wray & Grace, 2007; Trudgill, 2011; Hurford, 2011; see Chapters 4 and 5 for a fuller discussion).

Still, as detailed in Chapter 5, there are several pitfalls with this approach, especially when trying to infer causal relationships. Operationalising context is particularly difficult, and hard to parcel out from other factors, such as frequency (but see: Winter & Ardell, 2016⁸). Computational models provide another route of investigation: these offer a useful way for exploring the parameter space in a transparent and quantifiable manner (Irvine, Roberts & Kirby, 2014). Operationalising context is easier as it can be implemented as a parameter in a model. However, models are heavily reliant on the simplifying assumptions of the researcher, making it difficult to connect the results of the model to real world phenomena.

⁸Winter & Ardell (2016) show that, when controlling for *contextual diversity* (the number of different contexts a word appears in), word frequency does not predict the number of word senses. Importantly, the results of their statistical analyses are robust to operationalising contextual diversity using different sources (e.g., the number of different movies, the number of different Ngrams etc).

Experiments represent a natural middle ground by allowing us to operationalise context whilst testing predictions about causal relationships. As with other methods, experiments come with their own set of inherent limitations, such as circumscribed control over the parameter space when compared to computational models, and lacking in the richness of real linguistic data available for analysis with statistical techniques. However, as Chapters 2, 3 and 4 will demonstrate, experiments offer a powerful tool for investigating the relationship between context and language. For this thesis I will draw upon work in iterated learning (e.g., Kirby, Cornish & Smith, 2008) and communication games (e.g., Galantucci, 2005). These studies often use artificial languages (Reber, 1967; Saffran et al., 1996) as simplified abstractions for natural languages, and see under what conditions a language transitions from one state to another. This allows us to probe how individual-level biases, found in learning and using a language, contribute to the emergence of systematic structure. For operationalising and manipulating context, reference games offer a useful entry point (Olson, 1970; Sedivy, 2005; Frank & Goodman, 2012; Franke & Degen, 2016): here, manipulations to the *referential context* – the relationship between the intended referent and a set of possible alternative referents – have been previously shown to influence pragmatic reasoning (Franke & Degen, 2016) and the use of referring expressions (Konopka & Brown-Schmidt, 2014). Chapters 2 and 4 will further motivate these experimental paradigms.

1.6 Thesis road map

The thesis is structured as follows.

Chapter 2 reviews some of the key literature linking context to how language is learned, used and structured. It then presents Experiment 1, which builds upon previous work in iterated learning and communication games by manipulating the referential context, i.e., which dimensions of a referent are informative and uninformative for discriminating between other possible referents. The results of this experiment shed light on how short-term communicative interactions, where the goal is for the speaker to convey a specific referent to the hearer, shape and constrain the long-term emergence of linguistic structure as it is transmitted to future generations: that is, language systems gradually evolve to only specify information relevant for discrimination, resulting in markedly different systems of communication.

Experiments 2 and 3, presented in Chapter 3, take as their starting point the

observation that context motivates the loss of compositionality: the decrease in transparency between a form (e.g., *John kicked the bucket*) and its intended meaning (e.g., *John died*). Using a compositional language, where the subcomponents of a signal refer to specific features of a referent, these experiments see whether manipulations to the referential context increase the probability of maintaining or losing structure as the language is learned and used. While the results are inconclusive across these two experiments, there is tentative evidence that context can disrupt compositional structure, but only when interacting with other factors.

Chapter 4 investigates whether contextual predictability (the extent to which a speaker can estimate, and therefore exploit, the contextual information a hearer uses in interpreting an utterance) shapes the degree of signal autonomy (how reliant a signal is on contextual information for discriminating between possible referents). Experiment 4 uses an asymmetric communication game, where speakers and hearers are assigned fixed roles, to test for the effect of contextual predictability on signal autonomy by making two manipulations: (i) whether or not a speaker has access to contextual information; (ii) the consistency with which a particular dimension (e.g., shape) is relevant for discrimination across successive trials. The key finding is that, when the context is predictable, speakers organise languages to be less autonomous by combining linguistic signals with contextual information, whereas less predictable contexts result in speakers relying on strategies that promote more autonomous signals.

Chapter 5 addresses the advantages and limitations of applying statistical analyses to large-scale, cross-cultural datasets. In particular, the chapter highlights how the noisiness of the dataset, the borrowing of cultural traits, and the shared inheritance of cultures are all fundamental limitations on the explanatory power of such studies. Taken together, these three problems make correlational studies poorly equipped to test causal relationships, with the chapter concluding that experiments are better suited to answering questions relating to context and language structure.

The final chapter provides a summary of the work presented and offers some possible directions for future research.

Chapter 2

Languages adapt to their contextual niche

2.1 Introduction

Chapter 1 outlined a theoretical framework for how we might go about approaching the causal relationship between context and language structure, arguing that an experimental approach is best equipped to address such questions. The next three chapters investigate how context, learning and communication interact with one another in shaping the structure of an artificial language. The aim is to use experiments to tease apart the three pressures identified in Chapter 1: a *discrimination pressure* arising from a need to discriminate between objects in the world, the fact that languages need to generalise to new meanings and contexts (*generalisation pressure*), and that speakers and hearers need to align on a shared system of communication (*coordination pressure*).

Experiment 1 builds upon previous work in iterated learning and communication games to investigate how context links short-term language use with the long-term emergence of different types of language systems. To operationalise context we manipulate the *situational context*¹, defined as the physical environment in which an utterance is produced (Evans & Green, 2006), by changing which feature is informative (and uninformative) for discriminating between possible referents. This allows us to address the following questions: (i) To what extent does the situational context influence the conventional encoding of features in the linguistic system? (ii) How does the effect of the situational context work its way into the structure of language?

¹Also referred to as the *referential context* and *visual context* in other work (e.g., Konopka & Brown, 2014).

The results of this experiment show that languages gradually evolve to only encode information which is informative for the task of learning and using a language to discriminate between referents in context. The fact that different systems evolve for conveying the same set of referents demonstrates that context plays an important role in the emergence of structure in language.

2.2 Author contributions

The following section contains a paper which was co-authored with my supervisors, Simon Kirby and Kenny Smith, and published in *Language and Cognition*. The experiments were conceived during supervision meetings, with both co-authors contributing to the analysis and writing of the paper.

2.3 Winters, Kirby & Smith (2015): Languages adapt to their contextual niche

Languages adapt to their contextual niche

James Winters*, Simon Kirby, Kenny Smith

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

Keywords: cultural transmission; iterated learning; language evolution; context; communication games

Acknowledgements: We would like to thank Sean Roberts and two anonymous reviewers for feedback on an earlier draft of this manuscript. JW is funded by an AHRC studentship.

*Corresponding author: J.R.Winters@sms.ed.ac.uk

Abstract

It is well established that context plays a fundamental role in how we learn and use language. Here we explore how context links short-term language use with the long-term emergence of different types of language systems. Using an iterated learning model of cultural transmission, the current study experimentally investigates the role of the communicative situation in which an utterance is produced (SITUATIONAL CONTEXT) and how it influences the emergence of three types of linguistic systems: UNDERSPECIFIED languages (where only some dimensions of meaning are encoded linguistically), HOLISTIC systems (lacking systematic structure) and SYSTEMATIC languages (consisting of compound signals encoding both category-level and individuating dimensions of meaning). To do this, we set up a discrimination task in a communication game and manipulated whether the feature dimension shape was relevant or not in discriminating between two referents. The experimental languages gradually evolved to encode information relevant to the task of achieving communicative success, given the situational context in which they are learned and used, resulting in the emergence of different linguistic systems. These results suggest language systems adapt to their contextual niche over iterated learning.

1) Introduction

One of the fundamental axioms of modern cognitive-functional linguistics is that "[word] meaning is highly context-sensitive, and thus mutable" (Evans, 2005: 71). When interpreting a particular utterance, language users must not only rely on the meaning encoded in linguistic forms, but also on what they infer from contextual information. Such notions were explicitly acknowledged in the early work of Grice (1957), with a distinction being made between SIGNAL MEANING¹ and CONTEXTUAL MEANING (Evans & Green, 2006). Signal meaning refers to the senses stored in semantic memory, forming part of the user's linguistic knowledge. Contextual meaning is constructed *on-line* and constitutes an extension of the original signal meaning through an individual's inferential capacities (cf. Evans & Green, 2006; Hoefler, 2009: 6). Put simply: "[...] some meaning is encoded in linguistic forms and some is inferred" (Wedgwood, 2007: 652).

In this sense, context broadly refers to the set of premises used in interpreting an utterance, besides the information already specified in the signal meaning, and constitutes a psychological construct that

¹ We use *signal meaning* to refer to what Evans & Green (2006) refer to as *encyclopaedic meaning*.

comprises a subset of an individual's assumptions about the world (Sperber & Wilson, 1986: 15-16)². Consider the word MOLE. Besides referring to a small burrowing animal, MOLE can also denote a form of espionage, a type of birthmark and a unit in chemistry. Each of these senses are said to be stored in semantic memory, with their use and interpretation being governed by the very specific contexts in which they occur. Viewed in isolation words such as MOLE might be construed as communicatively dysfunctional. Yet, in context, it is typically easy to distinguish one sense from another. Having specific knowledge of the context thus enables a hearer to change their expectations regarding the intended meaning of a given word. In other words, when the context is known and informative, it necessarily decreases uncertainty (Piantadosi, Tily & Gibson, 2012).

As context is used as a resource to reduce uncertainty, it might alter our conception of how an optimal communication system should be structured (Zipf, 1949; Piantadosi, Tily & Gibson, 2012). Levinson (2000: 29), for instance, argues that our cognitive abilities favour communication systems which are skewed in their design towards hearer inference over speaker effort. Meanwhile, Pinker & Bloom (1990) note language exhibits design for communication because it allows for "minimising ambiguity *in context*" (pg. 713, emphasis added). Evidence for the role of context is also apparent in the way we structure our utterances, with syntax being sensitive to the wider discourse and the immediate communicative needs of interlocutors (Chafe, 1976; DuBois, 1987; Fery & Krifka, 2008). Furthermore, these immediate communicative needs can give rise to longer-term patterns: here, the way in which speakers pragmatically design utterances (INVITED INFERENCES, Traugott & Konig, 1991), as well as how hearers interpret utterances (CONTEXT-INDUCED INTERPRETATION, Heine, Claudi & Hunnemeyer, 1991), is posited to play a fundamental role in historical processes, such as grammaticalization (cf. Traugott & Trousdale, 2013).

There are a number of different kinds of context we could talk about in relation to a particular usage event (Evans & Green, 2006; Bach, 2012). Our present study is specifically focused on the SITUATIONAL CONTEXT: the immediate communicative environment in which an utterance is situated (Evans & Green, 2006: 221) and how it influences the distinctions a speaker needs to convey. In an experimental setting, situational context can be manipulated by tailoring both the types of stimuli and the way in which they are organised. For example, in a study examining how adjectives were used in referring expressions, Sedivy (2005) discovered that speakers were more likely to use an adjective when one object shared a feature dimension with another object (e.g., a blue cup and green cup), but not when the object belonged to a different category (e.g., a cup and a teddy bear). Similarly, Ferreira, Slevc & Rodgers (2005) found that when speakers were faced with conceptual ambiguities, such as

² This can refer to the wider *sentential context* in which a word is embedded as well as the *situational* and *interpersonal* contexts that make up the salient common ground, among others (see: Bach, 2012; Evans & Green, 2006: 221).

having to discriminate between two types of bat (the flying mammal), they would disambiguate on a relevant dimension (e.g., using *the small bat* in their utterance rather than just *the bat* when a large bat was also present in the context), whereas when speakers were presented with linguistic ambiguities (e.g., a baseball bat and an animal bat) they were less likely to engage in ambiguity avoidance.

If the situational context plays a fundamental role in how language is structured, then the general observation that *some meaning is encoded* and *some is inferred* leaves open the questions: (i) To what extent does the situational context influence the encoding of features in the linguistic system? (ii) How does the effect of the situational context work its way into the structure of language? To help answer these questions we investigate how situational context influences the emergence of linguistic systems. Using an artificial language paradigm, we experimentally simulate cultural transmission in a pair-based communication game setup (cf. Scott-Phillips & Kirby, 2010; Galantucci, Garrod & Roberts, 2012). Participants learn an artificial language which provides labels for a set of pictures, ‘meanings’ to be communicated. These stimuli vary on the dimension of shape, with each referent also having a unique, idiosyncratic element. After learning the language, participants play a series of communication games with their partner, taking turns to describe pictures for each other. We modified the situational context in which communication took place by manipulating whether the feature dimension of shape was relevant or not for a discrimination task: for example, some participants would encounter only situational contexts in which the objects to be discriminated during communication differed in shape, whereas others would be confronted with contexts in which the objects to be discriminated during communication were of the same shape. Finally, these pairs of participants were arranged into transmission chains (Kirby, Cornish & Smith, 2008; Scott-Phillips & Kirby, 2010; Thiesen-White, Kirby & Oberlander, 2011), such that the language produced during communication by the n th pair in a chain became the language that the $n+1$ th pair attempted to learn. This method allows us to investigate how the artificial languages change and evolve as they are adapted to meet the participants’ communicative needs and/or as they are passed from individual to individual via learning. We predict that languages in different types of situational context will adapt to become optimally structured as follows:

- When the feature dimension of shape always differs between pairs of referents which are to be discriminated, we predict that the languages will evolve to only encode shape in the linguistic signal, and become underspecified on all other dimensions.
- When the feature dimension of shape is always shared between pairs of referents which are to be discriminated, we predict that a holistic systems will emerge, in which each referent is associated with an idiosyncratic label that encodes that referent’s idiosyncratic feature;
- When the feature dimension of shape sometimes differs and is sometimes shared within pairs of referents, we predict that the languages will become systematically structured to encode

both the shape (via a category marker) and idiosyncratic features (via an individuating element of the signal).

1.1 Iterated Learning and Communication Games: A method for investigating the emergence and evolution of language

Language is not only a conveyer of cultural information, but is itself a socially learned and culturally transmitted system, with an individual's linguistic knowledge being the result of observing and reconstructing the linguistic behaviour of others (Kirby & Hurford, 2002). This process can be explored experimentally using ITERATED LEARNING: a cycle of continued production and induction where individual learners are exposed to a set of data, which they must then reproduce and pass on to the next generation of learners (Kirby, Cornish & Smith, 2008).

Using this method, researchers have demonstrated that cultural transmission can account for the emergence of some design features in language, including ARBITRARINESS (Thiesen-White, Kirby & Oberlander, 2011; Caldwell & Smith, 2012), REGULARITY (Reali & Griffiths, 2009; Smith & Wonnacott, 2010), DUALITY OF PATTERNING (Verhoef, 2012) and SYSTEMATIC COMPOSITIONAL STRUCTURE (Kirby *et al.*, 2008; Theisen-White *et al.*, 2011). Typically, a participant is trained on a target system (e.g., an artificial language) and then tested on their ability to reproduce what they have learned, with the test output being used as the training input for the next participant in a chain.

These studies show that cultural transmission can account for the emergence of structure in communication systems. In particular, communication systems adapt to constraints inherent in the learning process: domain-general limitations in our memory and processing capabilities (Christiansen & Chater, 2008) introduce a LEARNABILITY PRESSURE (Brighton, Kirby & Smith, 2005), meaning that languages that are difficult to learn tend not to be accurately reproduced, and therefore change. Recent work in this paradigm shows that the incorporation of situational context can change the extent to which the evolving language encodes certain features of referents. Silvey, Kirby & Smith (2014) show, using a transmission chain paradigm, that word meanings evolve to selectively preserve distinctions which are salient during word learning. Using a pseudo-communicative task, where participants needed to discriminate between a target meaning and a distractor meaning, the authors were able to manipulate which meaning dimensions (SHAPE, COLOUR and MOTION) were relevant and irrelevant in conveying the intended meaning. If a meaning dimension was backgrounded, in that it was not relevant in distinguishing between the target and distractor, then the languages evolved not to encode this particular meaning dimension. Instead, the languages converged on underspecified systems based on the relevant feature dimensions for discriminating between meanings.

However, language is not merely a task of passively remembering and reproducing a set of form-meaning pairings. Language is also a process of JOINT ACTION (Bratman, 1992; Clark, 1996; Croft, 2000): that is, language is fundamentally a social and interactional phenomenon, whereby the role of usage, communication and coordination are salient pressures on the system (also see: Tomasello, 2008; Bybee, 2010). Experimental communication games have been used to investigate the emergence of combinatorial (Galantucci, Kroos & Rhodes, 2010) and compositional (Selten & Warglein, 2007) structure, the emergence of arbitrary symbols from iconic signs (Garrod, Fay, Lee, Oberlander & MacLeod, 2007), and how common ground influences the extent to which a communication can become established in the first place (Scott-Phillips, Kirby & Ritchie, 2009).

Converging evidence from iterated learning and communication games point to both learning and communication as powerful forces in shaping the structure of language (Smith, Tamariz & Kirby, 2013; Fay & Ellison, 2013). With this in mind, the basic premise of the current experiment is to expand upon this work by: (a) adding a communicative element to the experimental setup of Silvey, Kirby & Smith (2014), and (b) manipulating the types of situational context.

1.2 The Problem of Linkage: Language Strategies and the emergence of language systems

Explaining how context works its way into the structure of language requires that we consider the PROBLEM OF LINKAGE (Kirby, 1999; Kirby, 2012). Rather than there being a straightforward link between our individual cognitive machinery and the features we observe in language, we are instead faced with an additional dynamical system: SOCIO-CULTURAL TRANSMISSION. Treating language as a COMPLEX ADAPTIVE SYSTEM (Beckner et al., 2009; Cornish, Tamariz & Kirby, 2009) solves this problem of linkage because we can consider how short-term LANGUAGE STRATEGIES (Evans & Green, 2006: 110) used in solving immediate communicative needs can give rise to LANGUAGE SYSTEMS through long-term patterns of learning and use (Bleys & Steels, 2009; Steels, 2012; Beuls & Steels, 2013).

The language strategy a speaker selects to enable a listener to identify their intended meaning is dependent not only on the referential information available, but also the context in which the utterance is situated. Take the relatively simple communicative situation in Figure 1: here, there are several language strategies that a language user could employ to convey the intended meaning. In context 1A, the intended meaning can easily be conveyed by using the label DOG as opposed to CAT. If, however, the situational context pairs the intended referent with another dog (as in context 1B), then it makes little sense to use the referential label of DOG, as the listener is very unlikely to be able to distinguish between the two referents on the basis of that label. Instead, other strategies must be employed, such as providing a unique identifier that is more specialised (DALMATIAN) or creating a compound signal (Ay, Flack & Krakauer, 2007) that has both specialised and generalised components (SPOTTED DOG).



Figure 1. Language strategies and example contexts. The green coloured boxes correspond to the intended referent. As we can see, DOG is a viable strategy for conveying the intended meaning in context A, but we need to either use a more specific label (DALMATIAN) or provide additional referential information alongside the generalised form (SPOTTED DOG) to convey the intended meaning in context B.

The current experiment explores how these short-term strategies of achieving communicative success in a situational context influence the emergence of different types of language systems. In particular, we focus on the evolution of three types of language systems: UNDERSPECIFIED, HOLISTIC and SYSTEMATIC. Underspecification captures the observation that languages abstract across referents by encoding some feature dimensions and ignoring others (Silvey, Kirby & Smith, 2014). Using the examples above, the word DOG is underspecified with respect to whether or not its referent is spotted or brown. Conversely, the labels DALMATIAN, POODLE, SIAMESE and TABBY are holistic, in that they

embody an arbitrary set of one-to-one mappings between signals and their meanings³: holistic signals serve the purpose of *individuation* (Lyons, 1977). Finally, in a systematic mapping between forms and meanings the signals share elements of form (unlike in a holistic mapping, where each signal is unrelated to the other signals) but are nonetheless one-to-one: systematic languages consist of compound signals (e.g., SPOTTED DOG), whereby part of the structure refers to a general-level category (e.g., DOG) and part of the structure refers to an individuating component (e.g., SPOTTED).

To test for the effect of situational context, we use a GUESSING GAME setup (cf. Steels, 2003; Silvey, Kirby & Smith, 2014): the task is to discriminate between pairings of a target object and a distractor object. In our case, possible referents are drawn from a set of images which vary in shape (see Figure 2 below). Manipulating these pairings gives us three experimental conditions based on: (a) whether the feature dimension of shape is relevant or not in discriminating between two referents, and (b) the extent to which stimuli pairings remain consistent over time with respect to the relevance of the feature dimension of shape.

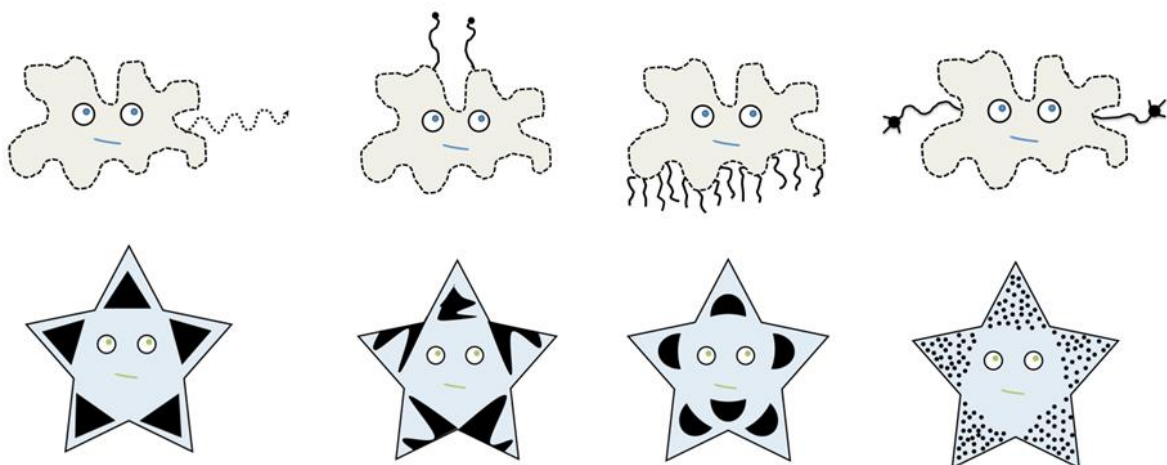


Figure 2. All eight meanings for the image stimulus set used in this experiment. Note that each individual image is comprised of two components: a basic-level of shape (star or blob) and a subordinate-level (a unique idiosyncratic feature).

In the SHAPE-DIFFERENT condition, pairings of target and distractor are constructed such that the feature dimension of shape is always relevant with respect to discrimination, i.e. target and distractor differ in shape. Since the two objects in such situational contexts have different shapes, then they can be discriminated merely by referring to shape. We therefore predict that the languages in the Shape-Different condition will evolve to become underspecified, specifying shape but not differentiating

³It should be noted that these holistic labels are also underspecified in respect to other possible signals (e.g., *dalmatian puppy*, *spotted dalmatian*). The important point to keep in mind is the relevance of dimension we are trying to signal. In this case, the signal *dog* is underspecified when compared with *dalmatian* and *poodle*. We could also highlight instances where *dog* is more specific than other signals, such as *animal*. Such notions are widely acknowledged in any linguistic theory that takes into account hierarchical relations between referents.

between the objects within a given shape category: such an underspecified system is functionally adequate for achieving communicative success in this situational context, and is highly learnable (Kirby, Cornish & Smith, 2008). Conversely, in the SHAPE-SAME condition target and distractor are always of the same shape – differing only on their idiosyncratic features. Consequently, the feature dimension of shape is always irrelevant in discrimination, and therefore does not need to be specified linguistically, with abstracting across referents of the same shape being communicatively dysfunctional for these situational contexts. We therefore predict that holistic systems will emerge in the Shape-Same condition, where each individual referent is associated with a unique label that maps onto its idiosyncratic feature. Lastly, for the MIXED condition we manipulated the predictability of situational contexts across trials: on some trials target and distractor share the same shape and on others they differ in shape. When encountering this mix of situational contexts, we hypothesise languages will become systematically structured, encoding in the linguistic signal both the basic-level of shape and individuating information of the idiosyncratic feature. Furthermore, we expect that the labels for the basic-level feature will become conventionalised earlier than those specifying the individuating information, with participants attempting to meet their immediate communicative needs on a piecemeal basis, through minimising effort and maximising communicative success; the quickest way to achieve this would be for participants to first align on conventional forms for two shapes (as this minimises effort and will ensure communicative success in contexts where shape is relevant in discrimination) followed by conventional forms for the eight idiosyncratic features (as these are needed to make these distinctions in contexts where shape is irrelevant in discrimination).

1.3 Ecologically-Sensitive, Learning Bias and Historically Contingent Accounts

Our prediction that manipulations to the situational context will bias the probability of one linguistic system emerging over another is consistent with a broader class of predictions that we will term ECOLOGICALLY-SENSITIVE accounts. Under this perspective, languages adapt to the structure of their niche in an analogous manner to that of biological organisms: just as environmental niches constrain and guide the evolution of species, so too are socio-cultural niches salient constraints on the types of languages that emerge (Lupyan & Dale, 2010). The ecologically-sensitive account is consistent with a range of observations including: that social structure patterns with differences in language structure (Wray & Grace, 2007; Lupyan & Dale, 2010); that word frequency is a product of the range of individuals and topics (Altmann, Pierrehumbert & Motter, 2011); that interactional constraints and conversational infrastructure lead to cultural convergence of linguistic form (Dingemanse, Torreira & Enfield, 2013); that objects and events in the world guide word learning discrimination (Ramscar, Yarlett, Dye, Denny & Thorpe, 2010); that word length patterns with the complexity of the meaning space (Lewis, Sugarman & Frank, 2014); that the structure of languages is shaped by the structure of meanings to be communicated (Perfors & Narravo, 2014).

These ecologically-sensitive accounts can be contrasted with two other theoretical perspectives that make different predictions about the relationship between the situational context and the emergence of linguistic systems. The first of these is the LEARNING BIAS approach. This makes the prediction that language structure is closely coupled to the prior expectations and biases of language learners (e.g. Griffiths & Kalish, 2007; Reali & Griffiths, 2009; Fedzechkina, Jaeger & Newport, 2012; Culbertson, Smolensky & Wilson, 2013; Culbertson & Adger, 2014). The learning bias approach can be further contrasted with what we term the HISTORICAL CONTINGENCY account, which holds that the types of systems that emerge are primarily constrained by random historical events, subtly biasing the language in one direction or another. When compared with the ecologically-sensitive and the learning bias accounts, a historical contingency prediction is that language structure is the result of lineage-specific outcomes (Lass, 1997), with “the current state of a linguistic system shaping and constraining future states” (Dunn *et al.* 2011: 1).

In their extreme incarnations, the learning bias and historical contingency accounts both predict that manipulating the situational context will have little effect on the types of systems that emerge in our experiment. For a learning bias account we would predict considerable convergence across all experimental conditions: there will be a globally-optimal solution in terms of a prior constraint (or set of constraints), with the languages then converging towards this prior. By contrast, the historical contingency account would predict a much higher degree of variation in the types of systems that eventually emerge, with the states of these systems being better predicted by individual variation and lineages than by either contextual or prior cognitive constraints.

2 Method

2.1 Participants

72 undergraduate and graduate students at the University of Edinburgh (42 female, median age 22) were recruited via the SAGE careers database and randomly assigned into 12 diffusion chains. Each chain consisted of a pair of initial participants who learned a random language, and two pairs of successive participants who learned the previous pair of participants' output language, making 3 generations in total. These chains were further subdivided into three experimental conditions (see §2.3).

2.2 Stimuli: Images and Target Language

Participants were asked to learn and then produce an alien language, consisting of lowercase labels paired with images. The images were drawn from a set of 8 possible pictures, which varied on the dimension of shape (4 blobs and 4 stars), with each individual image also having 1 unique, idiosyncratic subordinate element (see Figure 2).

The training language for the first participant pair in each chain was created as follows. From a set of vowels (a,e,i,o,u) and consonants (g,h,k,l,m,n,p,w) we randomly generated 9 CV syllables which we then used to randomly generate a set of 24 2-4 syllable words. These parameters ensured that there were 3 unique labels for every picture. Each chain was initialised with a different random language. The training language for later pairs of participants consisted of the language produced by the previous participant pair while communicating (see below).

2.3 Procedure: Training Phase and Communication Phase

At the start of the experiment, participants were told they would first have to learn and then communicate using an alien language. Participants completed the experiment in separate booths on networked computers. The experiment consisted of two main phases: a TRAINING PHASE and a COMMUNICATION PHASE. Before each phase began, participants were given detailed information on what that phase would involve and were explicitly told not to use English or any other language they knew during the experiment⁴. For the training phase, participants were trained separately, and it was only during the communication phase that they interacted (remotely, over the computer network).

2.3.1 Training Phase

In each training trial, the participant was presented with a label and two images, one of which was the target and one a distractor. The participant was told that the alien wanted them to pick which of the two images corresponded to the label. Once the participant had selected an image (by clicking on it using the mouse) they were told whether their choice was correct or incorrect, shown the label and target image for 2 seconds, and then instructed to retype the label before proceeding to the next trial. Both targets and distractors were presented in a random order within the following constraints: (i) the pairing of target and distractor varied based on the experimental condition (see §2.4 below for more details on the conditions); (ii) within each training block, each of the 8 meanings appeared three times as a target. The training phase of the experiment consisted of 4 such blocks, each of 24 trials; each block contained the same 24 training trails, with the order of these trials being randomly shuffled.

2.3.2 Communication Phase

During the communication phase of the experiment, participants took alternating turns as director and matcher:

- **DIRECTOR:** As directors, participants were presented with two images: a target and a distractor. Targets were highlighted with a green border. The director was prompted to type a

⁴ Compliance with the instructions was high – to our knowledge participants did not make use of English or any other language in the current experiment.

label that would best communicate the target to the matcher. The label was then sent to the matcher's computer.

- **MATCHER:** Participants were presented with the same two images as the director, with the label provided by the director appearing underneath. The matcher was then prompted to click on the image they thought corresponded to the label provided.

Following each trial, participants were given feedback as to whether or not the matcher had correctly identified the picture described by the director, followed by a display showing the image the director was referring to and the image the matcher selected. Target and distractor pairings were randomly generated within the constraints imposed by the experimental conditions (see §2.4 below), and communication trials were presented in random order. The communication phase consisted of 2 blocks, the length of each block varied depending on the experimental condition (see below).

2.4 Manipulating Context: Mixed, Shape-Same and Shape-Different Conditions

To test the role of context, a simple manipulation was made to the possible combinations of target and distractor images within a single trial during training and communication. This provides three experimental conditions. For the Shape-Same condition, participants only ever saw pairings of images that shared the same shape, but differed in their idiosyncratic element (see Figure 3A). In the Shape-Different condition, participants were exposed to pairings of images that differed in both their shape and idiosyncratic features (see Figure 3B). Participants in the Mixed condition encountered a mixture of image pairings: some image pairings shared the same shape but differed on their idiosyncratic features, whereas other image pairings differed on both their shape and idiosyncratic features (see Figure 3C).

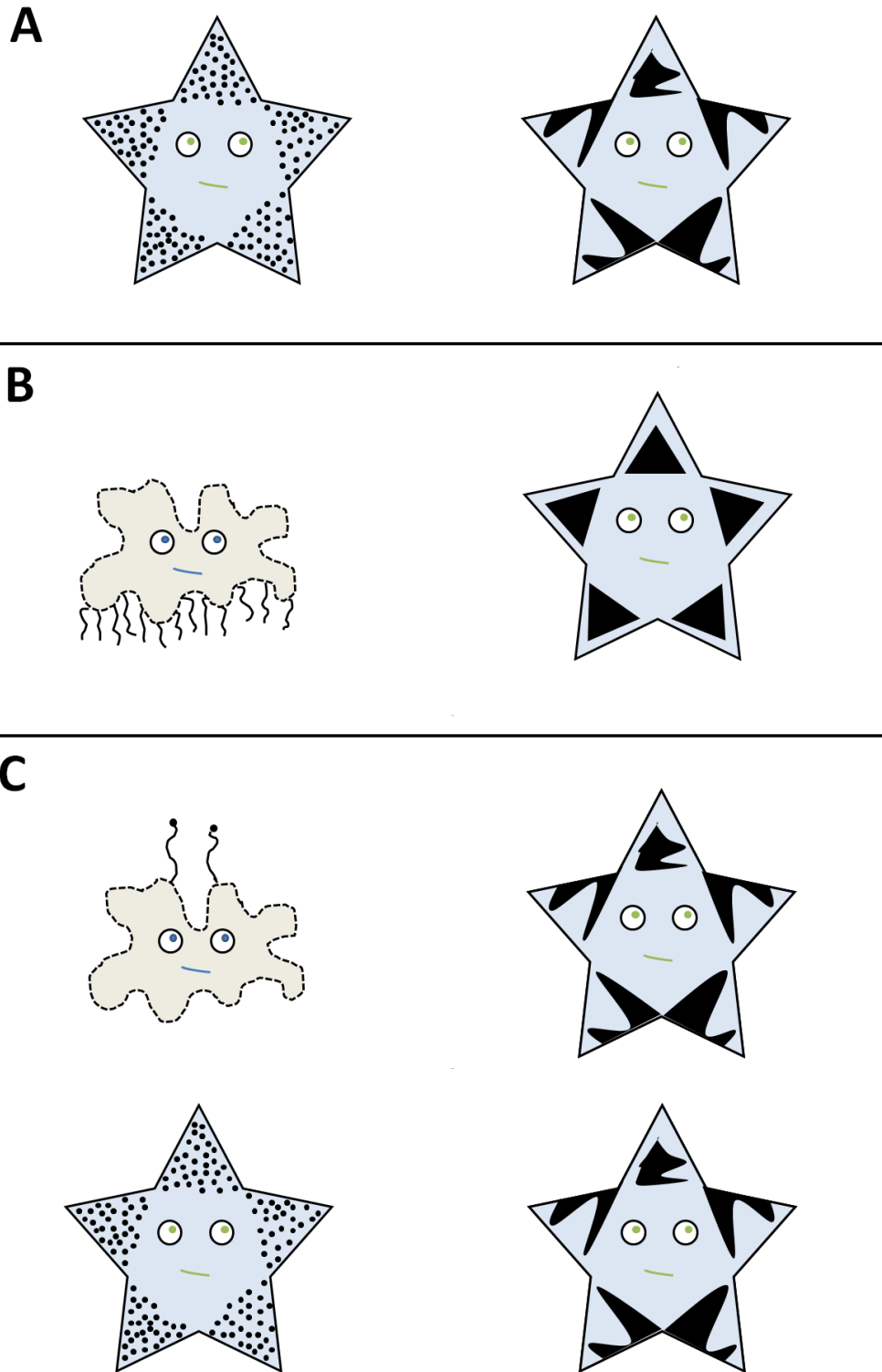


Figure 3. **3A** is an example of a pairing in the Shape-Same condition: here, participants only ever observe pairings that share the same shape. **3B** is an example of a pairing found in the Shape-Different condition; the two stimuli always differ in shape. **3C** shows an example of the pairings used in the Mixed condition: here, we get a mixture of stimuli that in some contexts differ in shape and in other contexts share the same shape.

In the Mixed condition, one communication block contained 56 trials, with 24 trials consisting of pairs of images that shared the same basic-level category but differed on subordinate-level features (24 trials exhausting all such possible pairings), whereas the remaining 32 trials differed on both their basic-level category and subordinate-level features (again, 32 trials covering all such possible pairings). To ensure that Shape-Different and Shape-Same conditions were comparable to the Mixed condition in the number of trials, we doubled up the possible combinations of images in the other two conditions, i.e., the Shape-Different condition involved 64 trials (32x2) per communication block and the Shape-Same condition involved 48 trials (24x2) per communication block; participants underwent two such blocks of communication.

2.5 Iteration

The labels produced by a pair of participants in the second block of the communication phase, and their associated target and distractor images, were used to construct the training language for the next pair of participants: we simply randomly sampled from the communicative output of generation n to produce the training language for generation $n+1$, (see Figure 4).

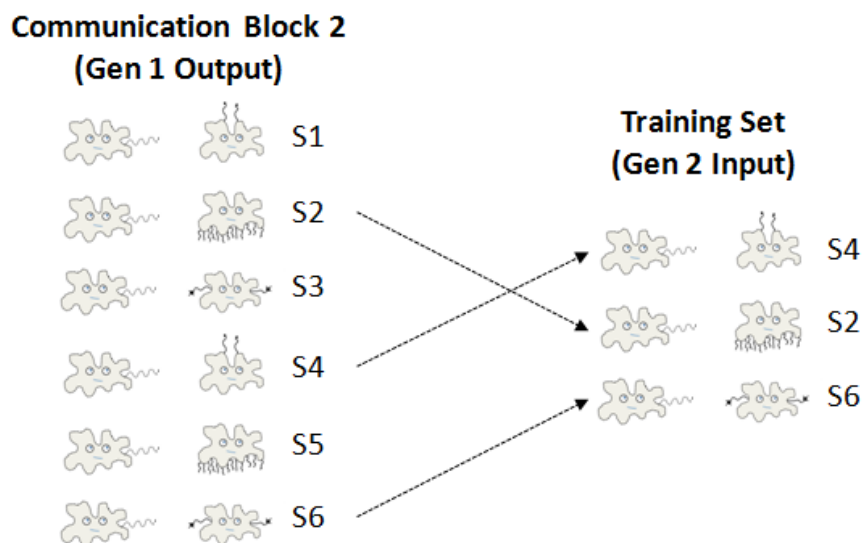


Figure 4. An example of the random selection process employed for a single meaning in the Shape-Same condition. Here, one target meaning is associated with 6 (possibly unique) signals during communicative testing. However, only three trials are required to construct a training block for the next generation: in order to generate this training block, we sample randomly from the appropriate contexts.

The random sampling process was constrained in the following ways. First, for all three conditions, we had a bottleneck on the number of signals that could be passed onto the next generation, i.e., for a single meaning we could only pass on three labels. As such, the number of signals transmitted from the final communication block of a generation stayed consistent between conditions, but the size of the sampling space differed slightly: Mixed (24/56 signals sampled), Shape-Different (24/64 signals

sampled), Shape-Same (24/48 signals sampled). Second, in the Mixed condition, the random selection process was additionally constrained so that a given stimuli would appear in at least one shape-same context and one shape-different context, and that there were an equal number (12) of shape-same and shape-different contexts in total. This meant that, in the Mixed condition, individual stimuli might appear in different ratios of shape-same and shape-different contexts. By contrast, the Shape-Same condition contained all possible pairings of target and distractor in training, and the Shape-Different condition had a subset of all possible contexts (24 out of 32 possible stimuli pairs).

2.6 Dependent Variables and Hypotheses

2.6.1 Measuring Communicative Success

To measure communicative success we simply recorded the number of successful interactions, where the matcher clicked on the target image. Given the differing trial numbers, the maximum success score differs across conditions: Shape-Different (128 points for two blocks of 64 interactions), Mixed (112 points for two blocks of 56 interactions) and Shape-Same (96 points for two blocks of 48 interactions). These maximum scores are converted into proportions to allow visual comparison between the three conditions, but the statistical analyses are conducted on the binary dependent variable.

2.6.2 Measuring Language Types: Difference Scores

In addition to conducting qualitative analyses of the languages that are produced during communication, we used the Normalised Levenshtein edit distance⁵ to provide objective measures for WITHIN-CATEGORY DIFFERENCE and BETWEEN-CATEGORY DIFFERENCE. To compute within-category difference for a given block, all labels associated with objects of a given category were compared with one another (i.e., all labels for the 4 blob-shaped images are paired with one another and given a total normalised Levenshtein edit distance, as were all labels for the 4 star-shaped images); the resulting pair of scores (a score for the blob-shaped category and a score for the star-shaped category) were then averaged to obtain a composite within-category difference score. Between-category difference was calculated for a given block by pairing all 4 labels for blobs with all 4 labels for star-shaped images at the same block and calculating average normalised Levenshtein distance.

These two difference scores provide us with an objective measure of language type. In particular, holistic, systematic and underspecified languages are discriminable on these scores, primarily the within-category difference scores. A holistic language only encodes the idiosyncratic feature of objects in the linguistic system – shape category distinctions are not encoded. As such, we should

⁵ The Normalised Levenshtein edit distance is calculated by taking the minimum number of edits (insertions, deletions, or substitutions of a single character) needed to transform one label into another, and then dividing by the length of the longer label.

expect the within-category and between-category differences to be similar. As a systematic language encodes both the shape category and the idiosyncratic element, systematic languages should exhibit smaller within-category difference scores than between-category difference scores, and should also exhibit lower within-category difference scores than holistic languages. For an underspecified language, we expect that only shape category information will be encoded, leading to substantial differences in within-category and between-category difference scores, with within-category scores being close to 0.

2.6.3 Measuring Uncertainty: Conditional Entropy

To further assist in quantifying the language types that emerge, we can calculate the degree of uncertainty in the system, which allows us to quantify the relationship between signals and their associated meaning. First, we need to operationalise two types of uncertainty about signal-meaning pairs. SIGNAL UNCERTAINTY arises from one-to-many pairings of meanings-to-signals (as in cases of synonymy in natural language). Conversely, MEANING UNCERTAINTY arises from one-to-many pairings of signals-to-meanings (as in cases of homonymy and polysemy in natural languages⁶). We predict that the languages in all three conditions will evolve over cultural transmission to lower their signal uncertainty: that is, as a system becomes more conventionalised, it is more likely to only have one signal for each meaning (cf. Reali & Griffiths, 2009). The Mixed and Shape-Same conditions are predicted to evolve toward a one-to-one mapping between signals and meanings (i.e. we should see eight signals for eight meanings in these conditions), leading to low meaning uncertainty. However, the Shape-Different condition is predicted to show higher levels of meaning uncertainty: the prediction is that these chains should involve one-to-many signal-meaning pairs, as an underspecified system leads to the same label being associated with multiple objects which share the relevant feature (here, shape).

To quantify signal uncertainty and meaning uncertainty we measure two aspects of the CONDITIONAL ENTROPY of the system. This gives us a measure of predictability that we can apply to both meaning uncertainty and signal uncertainty. $H(M|S)$ is the expected entropy (i.e. uncertainty) over meanings given a signal, and therefore captures meaning uncertainty,

$$H(M|S) = - \sum_{s \in S} P(s) \sum_{m \in M} P(m|s) \log P(m|s)$$

where the rightmost sum is simply the entropy over meanings given a particular signal $s \in S$. $P(m|s)$ is the probability that meaning, m is the intended meaning given that signal s has been produced. This entropy is weighted by a distribution $P(s)$ on signals. We can also reverse the position of signals and

⁶ We recognise there is a distinction between *ambiguity*, *vagueness* and *polysemy* in the lexical semantics and cognitive linguistics literature (cf. Tuggy, 1993; Geeraerts, 1993). For the sake of convenience, we use *meaning uncertainty* or *ambiguity* to simply refer to a one-to-many mapping of signals and meanings.

meanings in this equation to get the conditional entropy of $H(S|M)$, i.e., a measure of signal uncertainty:

$$H(S|M) = - \sum_{m \in M} P(m) \sum_{s \in S} P(s|m) \log P(s|m).$$

High $H(M|S)$ means that a signal is highly uninformative about the intended meaning (due to the signal having multiple meanings), whilst a high $H(S|M)$ means that a meaning is highly uninformative about the intended signal (due to the meaning having multiple signals).

While these measures capture relevant aspects of the structure of the evolving languages, they do not take context into account, and therefore do not capture the functional adequacy of the system for communication in context. To account for the contextual meaning we incorporate one last measure meaning uncertainty in context, $H(M|S, C)$,

$$H(M|S, C) = - \sum_{s, c \in S, C} P(s, c) \sum_{m \in M} P(m|s, c) \log P(m|s, c)$$

where the various sums are over signals and meanings GIVEN A CONTEXT. This measure captures the (potential) communicative utility of a system: we predict that the degree of in-context meaning uncertainty will decrease in all three conditions (the languages will be functionally adequate for conveying the correct/intended meaning), whereas meaning uncertainty (disregarding context) will differ across conditions depending on the emerging linguistic systems, as discussed above. As such, we are able to compare these two measures to provide an accurate account of how these types of systems are evolving over time, and whether or not they are adapting to their situational contexts.

2.6.4 Mixed Effects Model Overview

We used R (R Core Team, 2013) and *lme4* (Bates, Maechler & Bolker, 2012) to perform several separate linear mixed effects analyses based on the dependent variables of (a) communicative success, (b) within-category difference scores, (c) between-category difference scores, (d) $H(S|M)$, (e) $H(M|S)$ and (f) $H(M|S, C)$. For our independent variables, we entered condition (Mixed, Shape-Same and Shape-Different), generation and block as fixed effects with interactions. As random effects, we had random intercepts for chain and participant, as well as chain and participant random slopes for generation and block. Each of these models used the Mixed condition as a baseline category. Visual inspection of residual plots did not reveal any noticeable deviations from assumptions of normality or homoscedasticity. P-values were obtained using a MCMC sampling method (*pvals.fnc*) provided by the *languageR* package (Baayen, 2008).

2.6.4 Hypotheses

Here we recap and summarise our various hypotheses.

HYPOTHESIS ONE: Participants will increase their communicative success over successive blocks and generations.

HYPOTHESIS TWO: Languages in the Mixed condition will consistently evolve towards systematic category-marking systems.

HYPOTHESIS THREE: Languages in the Shape-Same condition will consistently evolve towards holistic systems.

HYPOTHESIS FOUR: Languages in the Shape-Different condition will consistently evolve towards underspecified systems.

HYPOTHESIS FIVE: The degree of signal uncertainty will decrease across all three conditions over successive blocks and generations.

HYPOTHESIS SIX: The Shape-Different condition is predicted to show higher levels of meaning uncertainty than the Mixed and Shape-Same conditions.

HYPOTHESIS SEVEN: The degree of meaning uncertainty in context will decrease across all three conditions.

3 Results

3.1 Qualitative Results: Languages

This section will provide an overview of a representative selection of languages observed in each of these three conditions. We contrast the initial starting language participants were trained on with very early systems at the start (generation 1, block 1 of communicative interaction) and at the end (generation 1, interaction block 2) of a single generation, as well as systems in the final generation of the chain (generation 3, interaction block 2).

Figure 5 shows an example from chain 1, from the Mixed condition. In generation 1, the labels for each individual referent tend to show some individuation: for instance, MUWUMUWU is only ever associated with one particular blob. However, even at this early stage, we start to see evidence that the labels are patterning systematically according to shape. For instance, the initial syllable MU is consistently associated with blob-shaped referents, and the template H*PA is associated with star-shaped referents. There is also some underspecification: HAPA, for instance, is used with all four stars (albeit at different frequencies). Words lengths also appear to differ systematically between shapes (although this strategy is not repeated in other chains). At the end of the first generation (block 2) a

few clear patterns emerge. First, the degree of heterogeneity has decreased in terms of the number of unique words and the number of unique syllables. Second, there is a higher degree of conventionality for each individual referent, as evident in some labels only ever appearing with one referent (e.g., MUHUMU and HEPA). Lastly, there is less underspecification across star-shaped referents – HAPA is now only associated with two stars. The language of the third generation extends these patterns of increased conventionality: each individual referent has a unique label that distinguishes it from other referents. Furthermore, these labels show systematic relations with one another: three of the blob-shaped images are distinguished from one another through varying the length of (partially) reduplicated syllables (MUWU, MUWUMU and MUWUMUWU). Meanwhile, all of the star-shaped images persist with the basic template of h*pa, and individual referents within this category differ only in the vowel of the first syllable. Finally, there is no underspecification by generation 3: as predicted, the language marks the basic-level category of shape as well as the individuating element. This observation supports our hypothesis that systematic structure will emerge in the Mixed condition, with languages first converging on conventionalised forms for shape followed by the idiosyncratic features.

















Input									Generation One Training data
	mewugu guwume hepaha	hawapame hapa pamehewe	pawumegu mumewuwe pamumu	wuwuwehe waweha pame	wupamu wahewu mewamu	wegu hehegume hemugume	muhapa hahewe hawegu	wawe hegu haguwu	
Output									Generation One Block One
	hegmug muwumuwu muwumuwu muwumuwu muwumuwu muwumuwu pumeh	wamuw wegmug uguhu umuwu muwehe muhewe umuhu	muwuhewe gumuhehe uguluh muguwu muguwu pumu muwu	muwegu huhu gumuwehe mewuhu hemehe muhuwe meheme	hapa hapa heweme hupa hupa hupa hupa	hapa hapa hapa hapa hapa hapa hapa	wumewe wemuwe humu mawe hapa hapa hapa	heme heme hapa hapa hapa hpa hopa	
Output									Generation One Block Two
	muwumuwu humuhu humuhu muhumu muhumu muhumu muhumu	umuhu umuhu muhumu muwu muhu muhu umuhu	muhumu muhumeh muhumu muhumu muhumu mawamawa mawa	muhuwe meheme h wumu humu meheme mawemew	hapa hapa hupa hapa hapa hepaa hapa	hapa hapa hapa hopa hopa hopa hepa	hepa hepa hepa hepa hepa hepa hepa	hopa hopa hopa hopa hopa hupa hopa	
Output									Generation Three Block Two
	muwumu muwumu muwumu muwumu muwumu muwumu muwumu	muwu muwu muwu muwu muwu muwu muwu	muwumuwu muwumuwu muwumuwu muwumuwu muwumuwu muwumuwu muwumuwu	meheme meheme meheme meheme meheme meheme meheme	hapa hapa hapa hapa hapa hapa hapa	hupa hupa hupa hupa hupa hupa hupa	hepa hepa hepa hepa hepa hepa hepa	hopa hopa hopa hopa hopa hopa hopa	

Figure 5. A table showing the initial training language and all of the signal-meaning pairs produced at generation 1 (communication block 1), generation 1 (communication block 2) and generation 3 (communication block 2) in chain 1 (Mixed condition). Each meaning appears with a collection of labels beneath it: this constitutes the combined output of a pair of participants in a particular generation.

In the first generation of the Shape-Same condition (Figure 6) we see some commonalities with the early stages of the Mixed condition: there are examples of conventionality (e.g., GIGI and ZARA) as well as diversity (e.g., the wide range of labels for the blob with antennae and the star with dots) in the labels used for the individual referents. By time we reach block 2 of the first generation there is almost a completely conventionalised system (in that the participants are aligned on a stable set of labels for each referent). Furthermore, unlike the Mixed condition, this conventionalised system tends to recycle holistic variants instead of introducing systematicity: while there are pockets of systematicity (e.g., KANAKU and NAKAKU), these are circumscribed when compared to the Mixed condition. Interestingly, at the third generation, the NAPAWE variant has been favoured over the NAKAKU variant, lending additional weight to the notion that the situational context is biasing the system AGAINST systematic structure. These observations provide support for our hypothesis that holistic languages evolve in the Shape-Same condition. However, we should note systematicity is tolerated to a certain extent, as is the case for the blob-shaped images (KAPA and KAPAPA and GUGU and GIGI).































Input									Generation One Training data	
	gimu lagihu kapa	moka gipawe wepa	gila gigi kala	mohupapa pakamo wegu	nagimuhu mogikamu hukahu	wemunahu hupa kawena	kalaka napawe nahu	kuhunamu nagimu kamuwemo		
Output									Generation One Block One	
	gugu gugu kapa kapa trala wepa	moka gaga moka gugu gugu	gigi gigi gigi gigi gigi	wepapa wepapa wepapa kapa wepapa kapapa	hukahu hukahu hakahu hukawe hukawe hokuwe	zara zara zara zara zara	zala kalaka napawe zala kalaka	bobu pada pada bobu bobu		
										Generation One Block Two
	kapa kapa kapa kapa kepa kapa	gaga gugu gugu gugu gugu	gigi gigi gigi gigi gigi	kapapa kapapa kapapa kapapa kapapa kapapa	kuluwe kuluwe kuluwe kanaku kanaku kanaku	zara zara zara zara zara	napawe napawe napawe napawe nakaku nakaku	bobu bobu bobu bobu bobu		
										
kapa kapa kapa kapa kapa	gugu gugu gugu gugu	gigi gigi gigi gigi	gugu kapapa kapapa kapapa kapapa	kanaku kanaku kanaku kanaku	zara zara zara zara	napawe napawe napawe napawe	bobu bobu bobu bobu			

Figure 6. A table showing the initial training language and all of the signal-meaning pairs produced at generation 1 (communication block 1), generation 1 (communication block 2) and generation 3 (communication block 2) in chain 6 (Shape-Same condition).

For the Shape-Different condition (Figure 7), we see that there is a high level of heterogeneity in both the labels used between and within the referents. There is, however, some clustering of syllable types (e.g., NO, GO, NI etc) and combinatorial patterns (e.g., PUGO, GOGO, PUMA) according to the basic-level category of shape. Interestingly, this diversity persists in the first generation (block 2), with less conventionality than that found in the Mixed and Shape-Same conditions. Still, there is an increase in conventional patterns, with forms becoming more predictable over time in both the number of syllables and the way in which they are arranged (e.g., ME and HE tend to disproportionately occur in the initial syllable position). The most noticeable difference between generation 1 and generation 3 is the collapse towards underspecification: we see high frequency forms for all blob-shaped referents (e.g., PUGU) and all star-shaped referents (e.g., HEHA). In addition to this loss of variation at the word level, variation also decreases at the syllable level (e.g., there are only four syllables for blob-shaped images: PU, PO, GU, and GO). The emergence of underspecified languages supports our hypothesis that languages in Shape-Same condition will evolve to abstract across the meaning dimension of shape.









									
Input	gugogo pupuheno pugono	higo nopuma gomahehi	henoni puhi puhanino	hemahiha puhigohi gugonini	hehima heha guhigono	nipuni hegoni nohihama	manohapu mahi gogohani	hamahi nogoma hagonihi	Generation One Training data
Output	nogogo nopuma pugoni nogogi pugoni pugoni pugoni nogonini	nopuma nogogi nopuma puma puma puma nugoni	pugoni pugoni nepuma nepuma pupunogi nopuma gogononi pugoni	gogonin gogonin nopuma puhani pugoni pupunogini pumagogi nugoni	hehami heha heha nehema nehema heramah nohema nemahena	manohi minaha nemaha hamahi nemahihi hamahi hamahi hamahi	hemaha hemaha hemaha hemaha henema nehema memeha memeha	manoha hamahi memaha nehaha hamahi hemeha hamahi hamahi	Generation One Block One
Output	nugino nugupo nogogo nupogo gogoni gogoni nugoni pugogo	puma puma puma puma nugoni pupino monogo pum	pupunoni pupunino pupuno pugoni pugini pupuno pogo punipu	punino pupino nopugo pupon nugopu punogo nopuma monogo	nehema nehema hemena hemaha namaha mahima mahima heha	hamahi nehema nehema nehema hamahi hemena hemena he	mehama mehama mehama mehama mahima heha hehama	maneha nehaha hamahi hehema meneha hemena harama hemana	Generation One Block Two
Output	pugu pugu pogo pogo pugogo pugogo pugu pugu	pugu pugo pugu pogo pogo pugo pugu pugu pugogo	pugu pugo pugu pogo pogo pugu pugu pugu	pugu pugogo pogo pogo pugu pugogo pugu	heha heha heha heha heha heha heha heha	heha heha mehima meheha heha heha heha meheha	heha heha heha heha heha heha heha mehima	heha heha heha heha heha heha hemaha mehima	Generation Three Block Two

Figure 7. A table showing the initial training language and all of the signal-meaning pairs produced at generation 1 (communication block 1), generation 1 (communication block 2) and generation 3 (communication block 2) in chain 12 (Shape-Different condition). Highlighted labels show underspecification.

It is important to note that all three conditions started off with a language that consists of randomly generated pairings of labels and meanings. Although the individual pairings differ between conditions, they do share an important structural characteristics: all initial languages have high levels of synonymy (three labels for each meaning). A consistent pattern shared across all three conditions is a shift from this system with many-to-one signal-meaning mappings to systems where we observe one-to-one and one-to-many mappings.

3.2 Communicative Success

Communicative Success scores tended to follow a similar trajectory in all three conditions (see Figure 8). Over successive blocks we observe a clear increase in the overall communicative success rate, leading to near-perfect communication by the end of generation 3. Analysis of the logistic mixed effects model revealed a significant main effect of Generation ($\beta = 1.13$, $SE = 0.19$, $z = 6.646$, $p < .001$) and Block ($\beta = 1.03$, $SE = 0.30$, $z = 3.399$, $p < .001$), but no effect of Condition and no other significant interactions ($p > .074$).

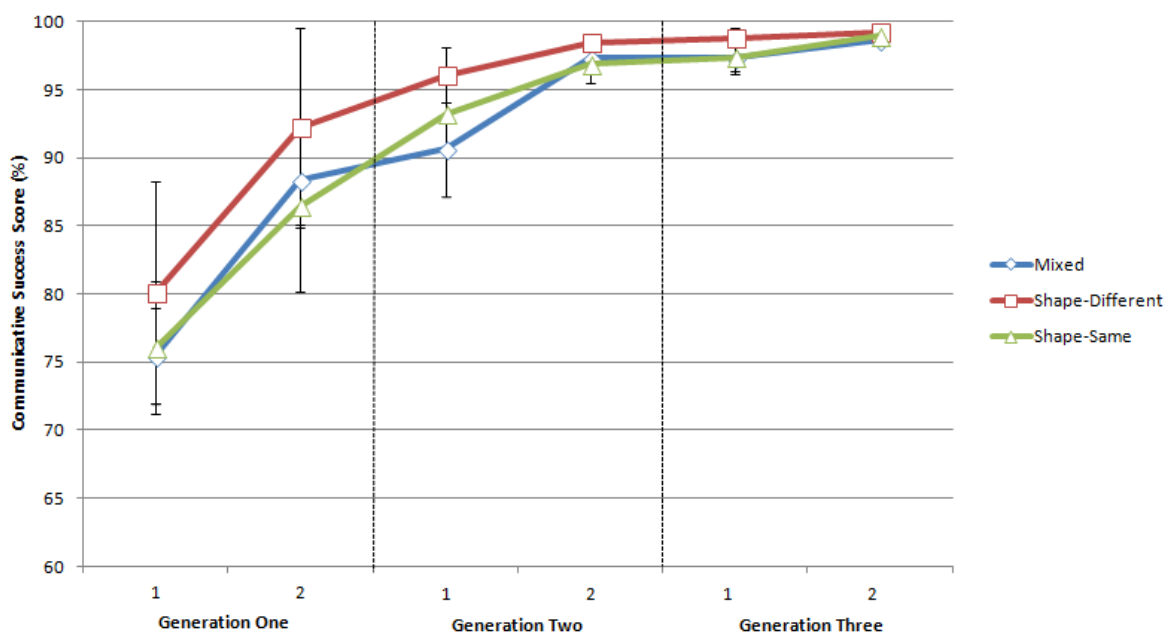


Figure 8. Average communicative success scores by generation (1-3), communication block, and condition. The vertical dotted lines represent the start of the next generation. Error bars represent the 95% confidence intervals.

These results show that, in all conditions, the languages are becoming increasingly effective at achieving communicative success through (a) repeated interactions between individual participant pairs and (b) across successive generations of participant pairs.

3.3 Difference Scores

Table 1 shows the idealised and observed (in the second block of generation 3) values for the within- and between-category difference measures. Figure 9 shows how these measures evolve over time. Our hypothesis that languages in the Mixed condition should evolve systematic category-marking and should therefore produce a within-category difference score of around 0.5 (characteristic of a system in which signals tend to be composed of a general category-marker and an individuating element) and a between-category difference score of 1 (distinctive labels used across categories). For the Shape-Same condition, we predicted the emergence of holistic languages, where each object is associated with a unique and distinctive label: this is characterised by high within- and between-category differences. As can be seen from Table 1, these predictions were borne out. For the Shape-Different condition, we predicted the emergence of systems that underspecified, using a single label for all objects sharing a shape, which would correspond to 0 within-category difference and a high between-category difference: as can be seen from the table, while this prediction was partially supported (within-category difference is lower than between-category difference), the within-category difference in this condition remains high – this is due to the slower conventionalisation seen in this condition, as highlighted in the qualitative analysis above (see also measures of signal uncertainty below).

Condition	Within-Category Difference		Between-Category Difference	
	Idealised	Observed	Idealised	Observed
Shape-Same	1	0.74 (SD = .05)	1	0.80 (SD = .07)
Mixed	0.5	0.47 (SD = .07)	1	0.88 (SD = .05)
Shape-Different	0	0.62 (SD = .10)	1	0.90 (SD = .08)

Table 1. The idealised (left-hand columns) and observed (right-hand columns) scores for within-category differences and between-category differences. Numbers in brackets indicate the bootstrapped standard deviation.

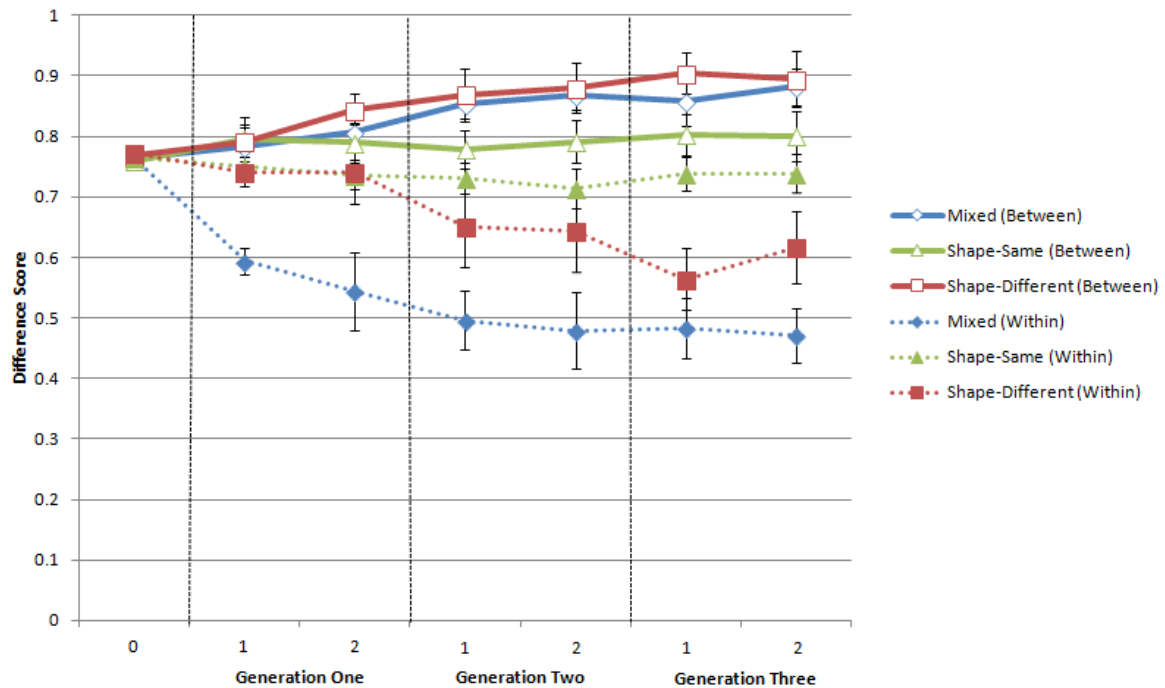


Figure 9. Between-category (solid lines) and within-category (dotted lines) difference scores (measured by the average normalised Levenshtein edit distance) over successive communication blocks for the Mixed (blue lines), Shape-Same (green lines) and Shape-Different conditions (red lines). Generation 0 gives values for the initial random language. Error bars indicate 95% confidence intervals.

Analysis of the mixed-effects model for Within-Category difference showed a significant effect of Generation ($\beta = -0.07$, $SE = 0.02$, $t(84) = -2.823$, $p < .001$), and a significant main effect of Shape-Same condition ($\beta = 0.20$, $SE = 0.04$, $t(84) = 5.043$, $p < .001$). There was one significant interaction for Shape-Same condition x Generation ($\beta = 0.07$, $SE = 0.03$, $t(84) = 2.266$, $p = .017$). All other main effects and associated interactions were non-significant ($p > .061$). These results partially support our predictions: within-category difference remains high in the Same-Shape condition, reflecting the development of labels which individuate within categories, and decreases in the other conditions; however, the Within-Category differences remain surprisingly high in the Shape-Different condition, where we predicted the emergence of a fully underspecified system, associated with Within-Category difference of 0.

Analysis of the model for Between-Category difference showed that only the main effect of Generation was significant ($\beta = 0.04$, $SE = 0.02$, $t(84) = 2.46$, $p < .001$), supporting the contention that Between-Category labels become increasingly distinct from one another over generations. All other main effects and associated interactions were non-significant ($p > .139$).

3.4 Conditional Entropy

3.4.1 Signal Uncertainty $H(S/M)$

For the conditional entropy of signals given meanings, $H(S|M)$, we observe a general decrease across all three conditions (see fig. 10). However, the decline in entropy for the Shape-Different condition appears to be less pronounced than that of the Mixed and Shape-Same conditions: as discussed above, within-category variation persists unexpectedly in this condition. For $H(S|M)$ the mixed-effects model contained significant results for the main effects of Generation ($\beta = -0.61$, $SE = 0.13$, $t(72) = -4.561$, $p < .001$), Block ($\beta = -0.38$, $SE = 0.09$, $t(72) = -4.366$, $p < .03$) and Shape-Different condition ($\beta = 0.62$, $SE = 0.31$, $t(72) = 2.011$, $p < .009$). There were no other significant main effects or interactions ($p > .259$).

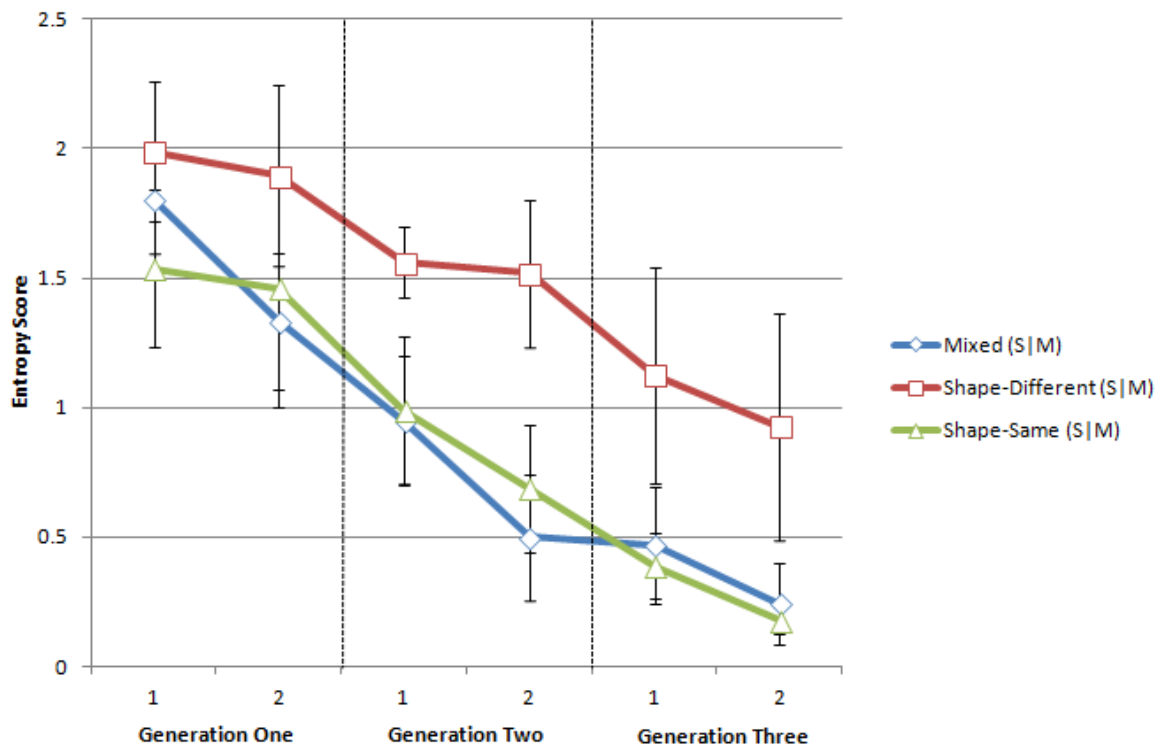


Figure 10. Degree of signal uncertainty, measured as $H(S|M)$, against Generation and Block. Higher entropy scores indicate a higher degree of signal uncertainty. The error bars indicate the 95% confidence intervals.

3.4.2 *Meaning Uncertainty $H(M|S)$*

Figure 11 plots the conditional entropy of meanings given signals, $H(M|S)$, against the number of blocks. As predicted, there is a clear difference between the conditions, with the Shape-Different condition showing a general increase in entropy in contrast to the Mixed and Shape-Same conditions, corresponding to the development of underspecified labels. For $H(M|S)$ the mixed-effects model contained significant results for the main effect of the Shape-Different condition ($\beta = 0.41$, $SE = 0.10$, $t(72) = 4.053$, $p < .001$). There was also a significant Shape-Different condition x Block interaction ($\beta = 0.32$, $SE = 0.09$, $t(72) = 3.424$, $p < .001$). There were no other significant main effects or interactions ($p > .265$).

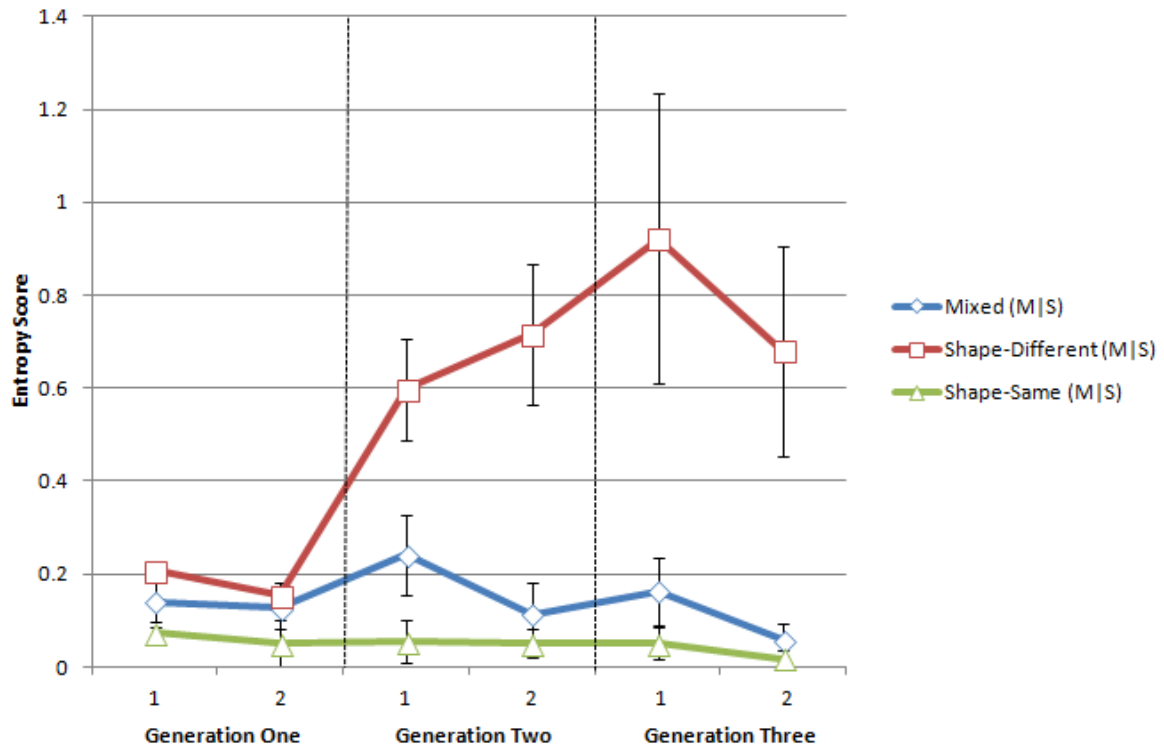


Figure 11. Degree of meaning uncertainty, measured as $H(M|S)$.

3.4.3 *Meaning Uncertainty of signals in context $H(M|S,C)$*

The conditional entropy of meanings given signals in context, $H(M|S,C)$, is shown in figure 12. In all three conditions we observe a decrease in entropy over time, with each of the conditions showing strikingly similar trajectories of change: as indicated by the communicative accuracy scores, the languages in all conditions evolve towards allowing optimal communication in context. For $H(M|S,C)$ the mixed-effects model contained significant results for the main effects of Generation ($\beta = -0.08$, $SE = 0.03$, $t(72) = -3.300$, $p < .001$) and Block ($\beta = -0.07$, $SE = 0.01$, $t(72) = -5.927$, $p < .001$). There were no other significant main effects or interactions ($p > .078$).

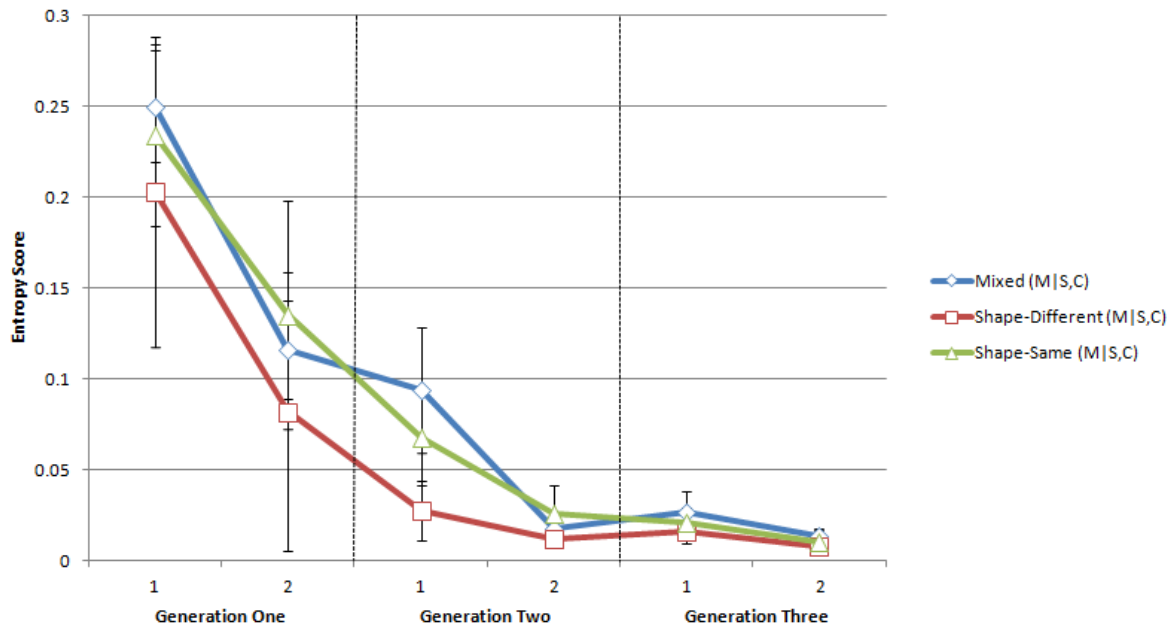


Figure 12. Meaning uncertainty of signals in context, measured as $H(M|S,C)$.

4 Discussion

Our findings support the general hypothesis that language structure adapts to the situational contexts in which it is learned and used. As we outlined in the introduction, some meaning is encoded and some meaning is inferred, with interactional short-term strategies of conveying the intended meaning feeding back into long-term, system-wide changes. In our experiment, languages gradually evolved to encode information relevant to the task of achieving communicative success in context, with different language systems evolving in each experimental condition. In the Shape-Same condition, where the dimension of shape was always the same for stimuli pairings, holistic systems of communication emerged, whilst in the Shape-Different condition, where the dimension of shape was always different for stimuli pairings, the system generalised and became underspecified (although unexpectedly variable: see discussion below). For the Mixed condition, which featured both Shape-Same and Shape-Different contexts, the systems that emerged were systematically structured: that is, both shape category and individual identity were encoded in the linguistic signal. These divergent systems arise given a very simple meaning space, through slight manipulations to the situational context.

Despite these inherent differences between the languages that emerged, all of the conditions showed: (a) an increased level of communicative success and (b) a reduction in in-context meaning uncertainty, $H(M|S,C)$. This observation suggests each condition produces languages that are functionally adequate for the task of achieving communicative success in context. The fact that different systems evolve for conveying the same set of meanings is important for how we view the role of context. Our explanation rests on the premise that languages are adapting to their niche, which in this case comprises the situational context, to become optimally structured.

Underspecified systems emerge in the Shape-Different condition because “when context is informative, any good communication system will leave out information already in the context” (Piantadosi, Tily & Gibson, 2012: 284). This lends weight to studies showing that participants are making use of pragmatic reasoning to convey information at the least cost given common knowledge and the task at hand (Frank & Goodman, 2012). These underspecified systems could be construed as being highly ambiguous when taken out of their communicative context. However, when we take into account the context in which the signals were used (as measured by the $H(M|S,C)$) then the apparent ambiguity is not counter-functional: that is, the system is perfectly adequate for achieving communicative success. When examined out of context, adapted communication systems can give the appearance of ambiguity, as Miller (1951: 111-2) noted: “Why do people tolerate such ambiguity? The answer is that they do not. There is nothing ambiguous about 'take' as it is used in everyday speech. The ambiguity appears only when we, quite arbitrarily, call isolated words the unit of meaning.”

While the amount of synonymy (as measured by $H(S|M)$) decreased over time across all conditions, the Shape-Different condition appeared to tolerate a higher level of synonymy than the other two conditions. One possible explanation is the way in which participants viewed the task. An initially diverse input could be construed as priming the participants to reproduce a diverse output. If the labels are easy enough to learn and reproduce, and they achieve the goal of successfully allowing the matcher to choose the correct image, then this variation may be tolerated for longer. This also partly explains why the Shape-Different condition deviates from its predicted within-category difference score: labels are not conventionally associated with any one particular meaning within a category. For instance, as discussed in the qualitative analysis (see Fig. 7), PUGU and POGO (which are quite distinct, with a normalised Levenshtein edit distance of 0.5) are not conventionally associated with any particular blob; instead, they pattern synonymously, with the two labels being optional forms for *any* blob-shaped image. This reflects a limitation of the difference measurement to distinguish between systematic languages and this kind of synonymy. However, these languages do have distinct profiles, as evidenced by the various entropy measurements.

It is also worth noting that not all chains in the Shape-Different condition converged on an underspecified system, with chain 11 evolving a holistic-like system. This mismatch with our predictions is perhaps due to the Shape-Different condition having more optionality provided by the situational context: that is, any of three hypothesised systems (Underspecified, Holistic, Systematic) are expressively adequate for conveying the intended meaning, although these systems differ in their parsimony in terms of memory and learning demands. This increases the probability that we will see more variation in the types of systems that evolve in the Shape-Different condition. Whereas

underspecified and, to a lesser extent, systematic category-marking languages are communicatively sub-optimal in the Shape-Same condition, the Shape-Different condition does not share such restrictions. A similar story applies when comparing the Mixed and Shape-Different conditions: neither holistic nor systematic category-marking languages are disfavoured for either condition, but an underspecified system would be problematic in the Mixed condition (as 43% of the contexts have images that share the same shape). Chain 11 thus serves as an important reminder of lineage-specificity, and how the historical properties of a particular system can bias future states.

For the Shape-Same condition, the chains consistently converge on holistic systems: that is, each individual stimulus has a unique label, with these labels being relatively distinct from one another. The decrease in $H(S|M)$ and $H(M|S)$ shows that the system is converging towards a one-to-one mapping of forms and meanings, whereas the high within-category difference scores show these signals are highly distinct from one another, and indeed more distinctive than those found in the other two conditions. Our rationale for the emergence of holistic systems in the Shape-Same condition is similar to that of the Shape-Different condition: where the situational context is informative, information will be left out of the linguistic system. In this instance, the context was informative through virtue of having the pairs of stimuli always sharing the same shape. This explains why systematicity is minimised in the Shape-Same condition: the linguistic system does not need to conventionally encode shape into the signal because context makes it irrelevant in discriminating between meanings. Instead, these languages specialise and become holistic, allowing them to meet the participants' communicative needs in context.

Even though the languages which emerge in the Shape-Same condition do reliably differ from those that evolve in the Mixed condition, through being more holistic, there is some evidence of systematicity in these chains. In chain 6, for instance, a language evolved in which two of the blob-shaped stimuli share similar labels (KAPA and KAPAPA) as do the other two blob-shaped stimuli (GUGU and GIGI). These pockets of correlations between word forms suggest a certain degree of systematicity is tolerated – albeit not to the same extent as that found in the Mixed and Shape-Different conditions. One explanation for this finding is that the situational context and communication are not the only factors shaping the system, with learnability pressures also acting on the structure of language (Kirby, Cornish & Smith, 2008).

Only in the Mixed condition do we consistently observe the emergence of systematic category-marking languages. The first line of evidence is that the observed within-category difference score lines up with our expected score (see Figure 11): this suggests part of the label is specifying shape and the other part is specifying the individuating component. While, as noted above, a difference score of approximately 0.5 is not necessarily indicative of systematic language structure, the $H(S|M)$ and

H(M|S) scores show that, by generation 3, the languages in the Mixed condition have low conditional entropy, showing that the form-meaning pairs embody one-to-one mappings.

A holistic language would be just as successful as conveying the correct meaning as a systematic language in the Mixed condition. So why do we see the emergence of systematic instead of holistic languages? Part of the reason rests on how these languages evolve in the early stages of their emergence: participants quickly establish a conventionalised specification of shape, before arriving upon conventionalised forms that encode the individuating elements. As a strategy, specifying shape information only requires participants to align on two signals, one that specifies star-shaped objects and one that specifies blob-shaped objects, which would allow them to successfully communicate on 57% of trials (those where discrimination only requires that shape information is conventionally encoded).

We can view this strategy as a negotiated exploration of the specification space during interaction, giving rise to a two-stage process: (i) THE CONVENTIONALISATION OF CATEGORY-MARKING FOR SHAPE; (ii) THE CONVENTIONALISATION OF INDIVIDUATING ELEMENTS. Supporting this contention of a two-stage process is the main effect of Generation for both the within-category difference scores and the conditional entropy of H(S|M): even though the within-category difference scores suggest systematic category-marking emerges by the end of generation one, the H(S|M) entropy is much higher in this initial generation than it is at later generations. The decrease in H(S|M) reflects the conventionalisation of individuating elements in the linguistic system - that is, there is less synonymy in later generations.

Another striking finding in the Mixed condition was the rate at which systematic category-marking emerged, within a single generation of participants. Part of the explanation could be in how the manipulation of context exerts a strong constraint for participants to quickly converge on conventional markers for shape. There are several reasons why the rapid evolution seen in this experiment might prove to be an exception, rather than a general tendency. First of all, there are only two possible dimensions that the language may encode: the basic-level category and the subordinate idiosyncratic component. There are also differences between the initial generation and successive generations (as mentioned above): namely, later generations show greater degrees of conventionalisation in their label usage.

If languages are adapting to their contextual niche, then what are the implications for the learning bias and historical contingency accounts? Even though our results are broadly consistent with the ecologically sensitive account, there is also evidence consistent with the learning bias (e.g., pockets of systematicity in the Shape-Same condition and the overall reduction of synonymy across all

conditions) and historical contingency (e.g., the emergence of a holistic language in chain 11 of the Shape-Different condition) accounts. It is likely that all these theoretical perspectives hold true to some extent, with the role of context being mediated by partially-competing motivations of prior learning biases and historical contingency. Such notions reflect the converging evidence that languages, and the way in which they are organised, “are better explained as stable engineering solutions satisfying multiple design constraints, reflecting both cultural-historical factors and the constraints of human cognition” (Evans & Levinson, 2009: 429).

5 Conclusion

We set out to investigate the role of situational context in the emergence of different types of linguistic systems that evolve through iterated learning. By manipulating the ways in which stimuli were paired with one another, we showed that situational context is an important factor in determining what is and is not encoded in the linguistic system. Our results offer a potential insight into how the situational context can bias the cultural evolution of language. The type and predictability of the situational contexts relate to how language users will employ certain communicative strategies for conveying the intended meaning, with the resulting language systems reflecting the contextual constraints in which they evolved.

One of the major findings in our experiment is that the types of linguistic systems that evolve are highly predictable based on their contextual constraints during communication. This interplay between short-term linguistic strategies for resolving communicative interactions, and the implication for language systems through long-term patterns of change, speaks to real-world processes such as grammaticalisation: the types of change we observe in languages show predictable patterns, as evident in the unidirectionality hypothesis (cf. Hopper & Traugott, 2003), but importantly these changes show how contextual constraints on the moment-to-moment communicative strategies deployed can have widespread ramifications on whole linguistic systems (Steels, 2012). Natural languages are subject to a larger and more diverse range of contexts, with a key future question being the extent to which our experimental results are generalisable to patterns observed in natural language systems.

6 References

Altmann, E.G., Pierrehumbert, J.B., & Motter, A.E. (2011). Niche as a Determinant of Word Fate in Online Groups. *PLoS ONE*, 6(5): e19009. doi:10.1371/journal.pone.0019009.

Ay, N., Flack, J.C. & Krakauer, D.C. (2007). Robustness and complexity co-constructed in multimodal signaling networks. *Philosophical Transactions of the Royal Society B*, 362: 441-447.

Baayen, R.H. (2008). *Analyzing Linguistic Data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Bach, K. (2012). Context Dependence (such as it is). In *The Continuum Companion to the Philosophy of Language*, M. Garcia-Carpintero and M. Kolbel, eds.

Bates, D.M., Maechler, M. & Bolker, B. (2012). *lme4: Linear mixed-effects models using Eigen and Eigenfaces*. R package version 0.999999-0.

Beckner, C., Blythe, R., Bybee, J., Christiansen, M.H., Croft, W., et al. (2009). Language Is a Complex Adaptive System: Position Paper. *Language Learning*, 59: 1–26.

Beuls, K., & Steels, L. (2013). Agent-based models of strategies for the emergence and evolution of grammatical agreement. *PLoS ONE* 8(3): e58960. doi: 10.1371/journal.pone.0058960.

Bleys, J. & Steels, L. (2009). Linguistic Selection of Language Strategies - A Case Study for Colour. *ECAL*, 2: 150-157.

Bratman, M. (1992). Shared cooperative activity. *The Philosophical Review*, 101: 327-341.

Brighton, H., Kirby, S., & Smith, K. (2005). Cultural selection for learnability: Three principles underlying the view that language adapts to be learnable. In Tallerman, M., editor, *Language Origins: Perspectives on Evolution*, chapter 13. Oxford: Oxford University Press.

Bybee, J. (2010). *Language, usage and cognition*. Cambridge: Cambridge University Press.

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

Chafe, W. (1976). Givenness, Contrastiveness, Definiteness, Subjects, Topics and Point of View. In *Subject and Topic*, ed. Charles N. Li, 25-56. New York: Academic Press.

Christiansen, M.H. & Chater, N.H. (2008). Language as shaped by the brain. *Behavioral and Brain Science*, 31(5): 489-508.

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

- Cornish, H., Tamariz, M. & Kirby, S. (2009). Complex adaptive systems and the origins of adaptive structure: what experiments can tell us. Special issue on Language as a Complex Adaptive System. *Language Learning*, 59(s1): 187-205.
- Culbertson J, & Adger D. (2014). Language learners privilege structured meaning over surface frequency. *PNAS*, 111(16): 5842-5847.
- Culbertson, J., Smolensky, P., & Wilson, C. (2013). Cognitive biases, linguistic universals, and constraint-based grammar learning. *Topics in Cognitive Science*, 5(3):392–424.
- Croft, W. (2000). *Explaining language change: An evolutionary approach*. Harlow, Essex: Longman.
- Dingemanse, M., Torreira, F., & Enfield, N.J. (2013). Is “Huh?” a Universal Word? Conversational Infrastructure and the Convergent Evolution of Linguistic Items. *PLoS ONE*, 8(11): e78273. doi:10.1371/journal.pone.0078273.
- Du Bois, J. W. (1987). The Discourse Basis of Ergativity. *Language*, 63(4): 805-855.
- Dunn, M., Greenhill, S.J., Levinson, S.C., & Gray, R.D. (2011). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, 473: 79-82.
- Evans, N. & Levinson, S.C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32(5): 429-492.
- Evans, V. & Green, M. (2006). *Cognitive linguistics: An introduction*. Edinburgh: Edinburgh University Press.
- Evans, V. (2005). The meaning of time: Polysemy, the lexicon and conceptual structure. *Journal of Linguistics*, 41: 33-75.
- Fay, N., & Ellison, T.M. (2013). The Cultural Evolution of Human Communication Systems in Different Sized Populations: Usability Trumps Learnability. *PLoS ONE*, 8(8): e71781.
- Fedzechkina, M., Jaeger, T.F., & Newport, E.L. (2012). Language learners restructure their input to facilitate efficient communication. *PNAS*, 109: 17897-17902.
- Ferreira, V., Slevc, L., & Rogers, E. (2005). How do speakers avoid ambiguous linguistic expressions? *Cognition*, 96(3): 263-284.

Fery, C. & Krifka, M. (2008). Information structure. Notional distinctions, ways of expression. *Unity and diversity of languages*, ed. Piet van Sterkenburg. Amsterdam: John Benjamins, 123-136.

Frank, M.C., and Goodman, N.D. (2012). Predicting Pragmatic Reasoning in Language Games. *Science*, 336: 998. doi: 10.1126/science.1218633.

Galantucci, B., Garrod, S., and Roberts, G. (2012). Experimental Semiotics. *Language and Linguistics Compass*, 6: 477–493. doi: 10.1002/lnc.

Galantucci, B., Kroos, C., and Rhodes, T. (2010). The effects of rapidity of fading on communication systems. *Interaction Studies*, 11: 100–111.

Garrod, S., & Galantucci, B. (2011). Experimental Semiotics: A Review. *Frontiers in Human Neuroscience*, 168 (6).

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.

Geeraerts, D. (1993). Vagueness's puzzles, polysemy's vagaries. *Cognitive Linguistics*, 4(3):223-272.

Gigerenzer, G. & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, 62: 451-82.

Grice, H. P. (1957). Meaning. *Philosophical Review*, 66: 377-388.

Griffiths, T.L., & Kalish, M. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science*, 31: 441-480.

Heine, B., Claudi, U., & Hünnemeyer, F. (1991). *Grammaticalization: A Conceptual Framework*. Chicago: University of Chicago Press.

Hoefler, S.H. (2009). *Modelling the role of pragmatic plasticity in the evolution of linguistic communication*. PhD Thesis, University of Edinburgh.

Hopper, P. J. & Traugott, E. C. (2003). *Grammaticalization*, Cambridge Textbooks in Linguistics: Cambridge University Press.

Kay, P. (1977). Language evolution and speech style. In: Blount, B.G. & Sanches, M. (Eds.), *Sociocultural dimensions of language change*. Academic Press: New York, pp. 21-33.

Kirby, S. (1999). *Function, Selection and Innateness: the Emergence of Language Universals*. Oxford: Oxford University Press.

Kirby, S. (2012). Language is an Adaptive System. The Role of Cultural Evolution in the Origins of Structure. In M. Tallerman & K. R. Gibson (Eds.), *The Oxford Handbook of Language Evolution*: 589-604. Oxford: Oxford University Press.

Kirby, S., Cornish, S. & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *PNAS*, 105: 10681-10686.

Kirby, S. & Hurford, J. (2002). The emergence of linguistic structure: An overview of the iterated learning model. In A. Cangelosi & D. Parisi (Eds.), *Simulating the Evolution of Language*. London: Springer Verlag, pp 121-148.

Kamide, Y., Altmann, G., & Haywood, S. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, 49(1): 133-156.

Levinson, S. (2000). *Presumptive meanings: The theory of generalised conversational implicature*. MIT Press.

Lass, R. (1997). *Historical linguistics and language change*. Cambridge University Press: Cambridge.

Lewis, M., Sugarman, E., & Frank, M. C. (2014). The structure of the lexicon reflects principles of communication. *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*.

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

Lyons, J. (1977). *Semantics*. Cambridge University Press: Cambridge.

Miller, G.A. (1951). *Language and Communication*. McGraw-Hill Book Company: New York.

Perfors, A. & Navarro, D.J. (2014). Language evolution can be shaped by the structure of the world. *Cognitive Science*, 38(4): 775-93.

Piantadosi, S.T., Tily, H. & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, 122: 280-291.

Pinker, S. & Bloom, P. (1990). Natural language and natural selection. *Behavioral and Brain Sciences*, 13: 707-784.

R Core Team. (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thrope, K. (2010). The effects of Feature-Label-Order and their implications for symbolic learning. *Cognitive Science*, 34: 909-957.

Real, F. & Griffiths, T.L. (2009). The evolution of linguistic frequency distributions: Relating regularisation to inductive biases through iterated learning. *Cognition*, 111: 317-328.

Scott-Phillips, T.C., Kirby, S., & Ritchie, G.R.S. (2009). Signalling signalhood and the emergence of communication. *Cognition*, 113(2): 226-33.

Scott-Phillips, T. C., & Kirby, S. (2010). Language evolution in the laboratory. *Trends in Cognitive Sciences*, 14(9), 411-417.

Sedivy, J. C. (2005). Evaluating explanations for referential context effects: Evidence for Gricean Mechanisms in Online Language Interpretation. In J.C. Trueswell & M.K. Tanenhaus (eds.): *Approaches to studying world-situated language use: Bridging the language as product and language as action traditions*, MIT Press, Cambridge, MA.

Selten, R., & Warglien, M. (2007). The emergence of simple languages in an experimental coordination game. *PNAS*, 104: 7361–7366.

Silvey, C., Kirby, S. & Smith, K. (2014). Word meanings evolve to selectively preserve relevant distinctions over cultural transmission. *Cognitive Science*. doi: 10.1111/cogs.12150.

Smith, K., Tamariz, M. & Kirby, S. (2013). Linguistic structure is an evolutionary trade-off between simplicity and expressivity. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 1348-1353). Austin, TX: Cognitive Science Society.

- Smith, K. & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition*, 116: 444-449.
- Sperber, D. & Wilson, D. (1995). *Relevance: communication and cognition*, Second Edition, Oxford/Cambridge: Blackwell Publishers.
- Steels, L. (2003). Evolving grounded communication for robots. *Trends in Cognitive Sciences*, 7(7): 308-312. doi: 10.1016/S1364-6613(03)00129-3.
- Steels, L. (2012). Self-organization and Selection in Cultural Language Evolution. In: Luc Steels, ed., *Experiments in Cultural Language Evolution*. John Benjamins: Amsterdam.
- Tomasello, M. (2008). *Origins of human communication*. Cambridge, Mass.: MIT Press.
- Thiesen-White, C., Kirby, S. & Oberlander, J. (2011). Integrating the horizontal and vertical cultural transmission of novel communication systems. In: *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, pp. 956-961.
- Traugott, E. & König, E. (1991). The semantics-pragmatics of grammaticalization revisited. In *Approaches to Grammaticalization*, eds., Elizabeth Traugott and Bernd Heine: 189-218.
- Traugott, E. & Trousdale, G. (2013). *Constructionalization and Constructional Changes*. Oxford: Oxford University Press.
- Tuggy, D. (1993). Ambiguity, polysemy and vagueness. *Cognitive Linguistics*, 4: 273-291.
- Verhoef, T. (2012). The origins of duality of patterning in artificial whistled languages. *Language and Cognition*, 4(4): 357-380.
- Wedgwood, D. (2007). Shared assumptions: Semantic minimalism and relevance theory. *Journal of Linguistics*, 43: 647-681.
- Wray, A. & Grace, G.W. (2007). The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua*, 117: 543-578.
- Zipf, G. (1949). *Human behavior and the principle of least effort*. New York: Addison-Wesley.

2.4 Conclusion

This chapter set out to investigate how manipulations to the situational context shape the emergence of distinct systems of communication. In particular, the bias introduced by the situational context, based on which feature is informative for discrimination, is amplified over repeated instances of learning and using a language. The experiment also provided proof-of-concept for systematically investigating the effects of context: by manipulating the situational context, we directly influence the frame of interpretation created in the minds of interlocutors, with inferences about form-meaning mappings being made in relation to which dimensions are informative and uninformative for discrimination.

More broadly, the results provide a stepping stone for thinking about language as a complex adaptive system, with language structure being the result of multiple, interacting pressures. In this experiment, the situational context imposed a strong discrimination pressure through providing a clear indication as to what is and is not informative for discrimination. However, this pressure interacts with two other motivating factors: learning and communication. For learning, a memory bottleneck is present, with participants at earlier generations being unable to remember all of the form-meaning mappings in their input data (generalisation pressure). This forces participants to generalise and use forms which are learnable *and* expressive. For communication, participants are trying to solve the immediate communicative problem, where the goal is to discriminate between possible referents, as well as the long-term problem of aligning on a shared system of communication (coordination pressure). Having generalisable forms, which are capable of discrimination in context, solves the coordination problem as a set of conventional form-meaning mappings can be established².

²All 95% CIs were bootstrapped by resampling the original data sample 10000 times using the *Hmisc package* (Harrell Jr et al., 2014). This allows us to approximate a CI for the sample statistic without making assumptions about the shape of the distribution (it is nonparametric).

Chapter 3

From *cup board* to *cupboard*: The loss of compositionality in linguistic systems

3.1 Introduction

Chapter 2 showed how manipulations to the referential context can result in radically different systems of communication based on what features of the meaning space are informative (and uninformative) for discrimination. As with all experiments, this general conclusion operates under a set of assumptions, which includes the initial set of form-meaning mappings an individual is exposed to (i.e., the training language). In the case of Experiment 1, as well as with previous experimental work in this area (e.g., Kirby, Cornish & Smith, 2008; for review, see: Tamariz & Kirby, 2016), the emphasis is on why a system transitions from an unstructured (e.g., a randomly generated language) to a structured (e.g., compositional language) set of signals. This is useful if, following Lindblom, MacNeilage & Studdert-Kennedy (1984: 187), our goal is to “Derive language from nonlanguage!”. However, this transition from unstructured to structured is only one possible set of experimental scenarios we can explore, and there are good reasons why we should consider other types of initial languages.

One such motivation is the patterns observed across historical timescales where changes take place in highly structured languages. If our goal is to model such changes, then it makes little sense to start from a position where the language is unstructured. For instance, imagine investigating word order change in English: in this case, it is useful to start with an initial language which has free word order (e.g., forms that alternate between Object-Verb (OV) and Verb-

Object (VO)), and see under what conditions it transitions to the more rigid VO. The second point to make is that experimental studies rarely focus on situations where structure is lost: that is, the focus is on how structure gets into language, and not on instances where structure is lost within the system. Yet historical linguistics is often concerned with questions that pertain to the loss of structure in individual form-meaning mappings.

Consider the English noun *cupboard*. Originally, *cupboard* consisted of two independent morphemes, *cup* and *board*, which combined to form a compound noun that meant a piece of wood used for displaying cups (and could be productively contrasted with other types of *board*, such as *washboard* and *cutting board*). Nowadays, *cupboard* refers to a closed storage area, having shifted in morphological status from a complex to a simplex form (Traugott & Trousdale, 2013: 22-23). The history of English is replete with similar examples, from the lexicalization of compound nouns, such as the aforementioned *cupboard* but also *neighbour* and *mincemeat* (Bussmann, 1996; Brinton & Traugott, 2005), to the emergence of idiomatic forms (e.g., *kick the bucket* and *pull strings*) (Bybee, 2010) and grammatical constructions (e.g., *be going to* and *let us*) (Hopper & Traugott, 2003).

A recurring theme across all of these examples is a *loss of compositionality*¹. Compositionality is where “the meaning of an expression is a function of the meaning of its parts and the way they are syntactically combined” (Partee, 1984: 281). From this point of view, a loss of compositionality simply refers to a mismatch between the individual meaning of the parts, and the interpretation of the whole sequence. Returning to our idiom example: *John kicked the bucket* can have a literal interpretation, whereby a person (*John*) performed an action (*kick-ing*) on an object (*a bucket*), or a non-literal interpretation (*John died*). Only in the second interpretation do we observe the loss of compositionality in that a hearer cannot use an underlying combinatorial rule to access the intended meaning of a speaker – the transparency between form and meaning is lost (Traugott & Trousdale, 2013).

This loss of compositionality will remain a key focus for the rest of the chapter. In the next section, we will consider the underlying mechanisms for the loss

¹An important distinction needs to be made between the loss of compositionality and the loss of analysability (Langacker, 1987; Bybee, 2010). *Analysability* refers to the extent to which a language user can identify internal parts in an expression (Bybee, 2010). Even though these measures of compositionality and analysability are correlated they are clearly separate from one another. For example, both the idiom *kick the bucket* and the preposition *beside* have lost their original compositional mapping, but only *beside* has lost its analysability (which originated from the phrase *by the side of*).

of compositionality, *reinterpretation* and *chunking*, and discuss claims that this change is principally motivated by the context. In particular, context can generate ambiguity in how a form maps onto a meaning, resulting in a reanalysis where compositionality is lost. This relationship between the loss of compositionality and context is investigated in two discrimination game experiments.

Experiment 2 takes as its starting point the observation that context motivates the loss of compositionality. To test this claim we expose participants to an initially compositional language for describing coloured shapes, manipulating which dimension (shape or colour) is relevant for discrimination in both learning and communication. The results suggest that manipulations to context do have an effect, albeit one which interacts with a bias to encode shape. One possible source of this preferential encoding of shape is the unbalanced number of values for the shape and colour dimensions (i.e., there were more colours than shapes). In light of this, Experiment 3 follows up on these findings in two ways. First, we increase the number of participants, and second we control for the confound by introducing an additional manipulation where participants were either exposed to a Colour-Skewed (4 colours, 3 shapes) or a Shape-Skewed (3 colours, 4 shapes) meaning space. Taken together, these two experiments provide tentative evidence that manipulations to context can disrupt a compositional mapping, but only when interacting with other factors (e.g., a skew in the number of dimensions). The chapter will close by discussing some problems with the experimental design as well as highlighting potential directions for future work.

3.1.1 Loss of compositionality: reuse, chunking, and context

Considerable work has gone into elucidating the mechanisms underpinning the loss of compositionality (Estill & Kemper, 1982; Gibbs & Gonzales, 1985; Langacker, 1987; Millikan, 2001; Wray, 2002; Hopper & Traugott, 2003; Wray & Grace, 2007; Bybee, 2010; Cruse, 2011; Traugott & Trousdale, 2013). For the purposes of this chapter, we shall reduce these to just two, *reinterpretation* and *chunking*.

The first mechanism, *reinterpretation*, subsumes other explanatory mechanisms of semantic innovation, such as *analogy* (Hopper & Traugott, 2003), *categorization* (Bybee, 2010), and *metaphor* (Smith & Hoefler, 2014), under one general definition: an inferential innovation where a pre-existing form maps onto a new interpretation. Reinterpretation therefore focuses on the functional side of the loss of compositionality: it is concerned with how an original meaning (e.g., cup

+ board: a board on which cups are placed) comes to have a new interpretation (e.g., cupboard: a general purpose storage unit). Importantly, reinterpretation is viewed as the principle mechanism for semantic change, covering phenomena such as generalisation (e.g., *dog*: restricted → general), specialisation (e.g., *hound*: general → restricted), metaphor (e.g., *back*: a body part → a spatial meaning) and metonymy (e.g., *blue collar*: workers who wear blue shirts → a particular type of work) (Fortson, 2004).

As the lexical examples illustrate, reinterpretation need not be restricted to compositional constructions, and is a general mechanism involved in semantic innovation and change. For reinterpretation to apply to compositional constructions we need our second mechanism, *chunking*: a way of organising memory that allows us to take a set of sequential units and group them into larger units (Miller, 1956; Newell, 1990; Bybee, 2010). A non-linguistic example of chunking is number memorisation: the sequence 4-9-7-5-2-5 is easier to store in memory if we chunk the sequence into 497-525. In language, chunking allows for a repeated sequence of words or morphemes to be packaged together as a single unit, and is argued to be the primary mechanism behind the formation of constructions and constituent structure (Bybee, 2002; 2010). As a standalone mechanism, chunking is not sufficient to explain the loss of compositionality, as demonstrated by collocations: *knife and fork* is a chunked unit, in that it can be accessed as a single unit when compared with the dispreferred *fork and knife*, but importantly the meaning has not changed – *knife and fork* and *fork and knife* both refer to the same two types of cutlery. It is only when reinterpretation and chunking act together do we observe a loss of compositionality through *reanalysis*: “a mechanism which changes the underlying structure of a syntactic pattern and which does not involve any immediate or intrinsic modification of its surface structure” (Harris & Campbell, 1995: 61). To summarise: as a sequence of words or morphemes are used together they are reanalysed as a single, cohesive unit which, when combined with a shift in the underlying meaning, results in the creation of a new form-meaning mapping (also see *form-meaning reanalysis*: Croft, 2000).

Why do some constructions undergo a loss of compositionality whereas others do not? What is special about *cupboard* when compared with *cutting board* or *bookshelf*? As with many aspects of language change, historical contingency is a key factor in shaping the trajectories of change, and its role should not be downplayed (Lass, 1997). Still, linguists have identified some motivations for the loss of compositionality (Bybee, 2010; Traugott & Trousdale, 2013). Frequency effects (Bybee, 1985), and in particular relative frequency (Hay, 2001; 2002), are often cited as crucial motivating factors. Relative frequency refers to the

frequency of the base word (e.g., *mortal*) when compared with a derived form (e.g., *immortal*). Hay (2001; 2002) observed that morphologically derived forms which are more frequent than their bases (e.g., *abatement* is more frequent than *abase*) tend to be less compositional than complex forms which are less frequent than their base (e.g., *top* is more frequent than *topless*).

Relative frequency is clearly part of the story, but as Bybee (2010: 48) explains, “Semantic and pragmatic shifts that reduce compositionality are aided by frequency or repetition, but their source is in the contexts in which the complex unit is used”. What Bybee is referring to is that the loss of compositionality ultimately stems from inferences made by language users in particular contexts of learning and use. It has long been appreciated that context is used to enrich interpretation of utterances (*Principle of Contextuality*: Frege, 1884; Wittgenstein, 1921; Grice, 1957). Consider the sentence *To bank money at the bank on the river bank*: here, the word *bank* takes on three distinct meanings (a verb meaning to store, a financial institution, and a type of geological formation), each of which is framed by the internal linguistic context in which it is used. Out of context, *bank* is ambiguous with respect to these possible interpretations (even if there is a more prototypical meaning; Rosch & Mervis, 1975; Gärdenfors, 2000; Ramscar & Port, 2015). Context can also generate ambiguity. For example, *the man with the binoculars* has the unambiguous interpretation of a man who is in possession of binoculars, but when placed in a linguistic context, such as *the boy saw the man with the binoculars*, we get a two-way ambiguous construction: either the boy could be using the binoculars and saw a man *or* the man has binoculars and a boy saw him (Berwick et al., 2011).

Context is therefore tied to uncertainty in interpretation – and in some cases this uncertainty in interpretation becomes part of the conventional meaning through repeated inferences (see Heine, Claudi & Hünemeyer, 1991: *context-induced reinterpretation* and Traugott & König, 1991: *invited inferences*). Linguists normally distinguish between two types of context involved in this change: onset and isolation contexts. An *onset context* is where the context introduces the possibility of a new interpretation. To unpack this into a series of incremental steps, consider the loss of compositionality in the *be going to* construction. In Shakespeare’s English, *be going to* was unambiguously a purpose clause and could be interpreted as a subject travelling to a location to do something, yet due to the semantic generality of *go* the construction found itself in contexts where there was the potential for an ambiguous interpretation between a motion event and future intent (Bybee, 2003; Bybee, 2010; Traugott & Trousdale, 2013). The sentence, *I am going to marry Maria*, is an example where travelling is inherent

to the meaning, but the fact that it has not happened yet implies future intent. Repetitive use of *be going to* in these contexts leads to a chunking of the phrase, subsequently triggering a reanalysis along the lines of: [I am going [to marry Maria]] → [I [am going to] marry Maria]. This reanalysis now supports the inference that future is an inherent part of the meaning. *Isolating contexts* are distinct from onset contexts in that this new interpretation is used in an unambiguous way, i.e., the future meaning of *be going to* is unmasked due it being extended to contexts that were previously unavailable. The sentences *I am going to go and marry Maria* or *the leaves are going to fall off the tree* are isolating contexts for *be going to* as it results in an unambiguous interpretation: *be going to* must refer to an event in the future that is interchangeable with other future constructions (the leaves *will* fall off the tree) but not motion constructions (*the leaves are travelling to fall off the tree) (Bybee, 2003).

Under this perspective uncertainty in interpretation is viewed as the engine of change in the loss of compositionality. But what triggers this uncertainty are inferences made in particular contexts. The first step in this process is that an initially compositional construction is chunked and used in an onset context, generating the possibility of a new interpretation. In the second step, this new interpretation becomes associated with a chunked construction, and is then used in an isolating context: this severs the link between the original form-meaning mapping by allowing for an unambiguous interpretation of the new meaning.

3.1.2 Modelling the loss of compositionality

The general hypothesis is that manipulations to context can result in the loss or maintenance of compositionality. More specifically, context can increase or decrease uncertainty in interpretation, with an increase in uncertainty being causally related to a decrease in compositionality. If the initial mapping is compositional, and there is a high uncertainty in interpretation, then the probability of the hearer recovering this compositional mapping decreases – and thus requires an alternative interpretation as to how the signal maps onto the meaning.

How can this general hypothesis be investigated experimentally? The first step is to introduce an initially compositional language into the set up. This can be achieved using an artificial language paradigm where a set of signal forms are mapped onto a set of multidimensional meanings (see Chapter 2). The meanings used in this experiment will consist of images that vary on two dimensions: colour and shape. The signals are structured such that the initial CV syllable refers to colour and the final CVC syllable refers to shape (see 3.2.1 for more information on

the method). This allows participants to entertain three hypotheses about how a particular form maps onto a meaning. The first is that subcomponents of the form map onto the individual features of a meaning (compositional mapping). The second is that the whole form maps onto one feature (underspecified mapping). The third is that the whole form maps onto a single meaning (holistic mapping) (see Figure 3.1).

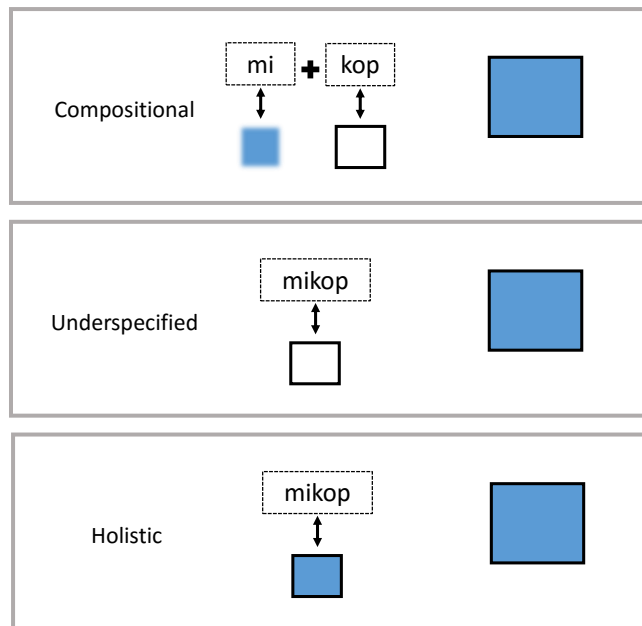


Figure 3.1: An example of three possible form-meaning mappings for the same referent (a blue square). From top to bottom: Compositional (where *mi* maps onto blue and *kop* maps onto square), Underspecified (where *mikop* maps onto square), and Holistic (where *mikop* maps onto blue square).

Which of these mappings is selected depends on the constraints participants bring to the task of learning and using a language. Two of these constraints are the domain-general cognitive bias of *simplicity* and the communication-specific bias of *expressivity* (Chapter 1; Kirby et al., 2015; Culbertson & Kirby, 2015). A simplicity bias makes the system more compressible, i.e., the description length of the system is shorter than a list of signal-meaning mappings, whereas the pressure from expressivity requires that the system is communicatively functional, i.e., capable of identifying the intended meaning in context. Languages must reach a tradeoff between these two constraints. How this tradeoff is reached is dependent on the way these constraints interact with the context and task.

All else being equal, an underspecified mapping is the simplest of the sys-

tems mentioned, as it can be described by a more compressible rule than either a compositional or holistic mapping. With this simplicity comes a decrease in expressivity: an underspecified mapping is only functional insofar as a particular feature remains relevant to discrimination (i.e., it is context-dependent). By contrast, a set of holistic mappings is the least compressible of all systems; a description of the system requires listing every signal and the meaning it maps onto (i.e., description length is equal to list length). Holistic mappings are highly expressive if discrimination takes place between meanings that are already known. However, a holistic language is costly for two reasons. First, it requires remembering an increasing number of unique form-meaning mappings, which over time creates a burden on memory (Kirby, Cornish & Smith, 2008; Cornish, 2010; Pleyer & Winters, 2014). Second, in a context consisting of two or more new meanings, discrimination becomes difficult (as, without any additional inferential clues, the new signal could refer to any of the new meanings; see Chapter 4 for a fuller discussion). Compositional mappings sit in-between underspecified (most compressible) and holistic (least compressible) in terms of simplicity: this allows compositional systems to be generalisable (i.e., it can convey new meanings) without a decrease in expressivity (i.e., disambiguation relies less on the external context) (Kemp & Regier, 2012; Kirby et al., 2015).

The only way an underspecified mapping is going to be favoured is if we amplify the bias for simpler form-meaning mappings. One way of achieving this is to manipulate what is and is not relevant for discrimination. We already established in Chapter 2 that manipulations to the referential context can shape the types of systems that emerge (also see: Silvey, Kirby & Smith, 2014). As in previous work, the context in this study is manipulated across conditions to make a particular dimension (e.g., shape) more relevant than another dimension (e.g., colour) for discriminating between meanings. When there is a discrimination pressure to only convey one dimension, we predict that this increases the probability of transitioning from a compositional to underspecified system. This gives us three conditions based on manipulations to the referential context. For the Shape-Different condition, trials are constructed so that shape is always the relevant feature for discriminating between a target and a distractor (which share the same colour). Conversely, the Colour-Different condition only ever has trials whereby colour is the relevant feature, with the target and distractor always sharing the same shape. Lastly, the Both-Different condition consists of trials where both shape and colour are (potentially) relevant for discrimination (that is, the target and distractor always differ on both colour and shape).

Leveraging the referential context in this way means that the optimal system

in all three conditions is underspecified: it is expressive (i.e., a capable of discriminating between meanings in a context) and it is compressible (i.e., the system can be described using a simple rule). For Shape-Different and Colour-Different conditions, the discrimination pressure directs the attention of participants toward the relevant feature and backgrounds the irrelevant feature – increasing uncertainty in interpretation as the probability of recovering the compositional mapping decreases. For Both-Different, the context cues both colour and shape – that is, these dimensions have equal salience in terms of discrimination. Therefore, our hypothesis is that uncertainty in interpretation is lower in Both-Different: participants have a higher probability of interpreting a compositional mapping as both dimensions are (potentially) relevant for successful discrimination.

The experimental set up also needs to control the amount of data participants are exposed to in training. By backgrounding a dimension, as is the case for Shape-Different and Colour-Different, we increase the bias against a compositional mapping due to one dimension being irrelevant for discrimination. Still, even though we predict an effect of context, it is unlikely this is enough to overcome the bias imposed by the full set of compositional mappings. This bias from the set allows the learner to easily recover the compositional system: that is, the signal-meaning data overwhelmingly favours a compositional mapping, with there being systematic correlations between the forms and the meanings. Training participants on a full set is also problematic for methodological reasons: it is hard to discern whether participants learned a compositional rule or simply memorised all of the form-meaning mappings.

To get around these issues we incorporate a *generalisation pressure* (see Kirby, Cornish & Smith, 2008; Cornish, 2010) into the set up: this involves training participants on a subset of all possible signal-meaning pairs and then placing them in a communication task where they need to convey the full set of meanings. By looking at how participants generalise, we can investigate whether they are learning the compositional rule and using novel combinations to express unseen meanings, or whether they reuse pre-existing forms to generalise across multiple meanings (underspecification). As such, the learning task entails that participants infer the signal-meaning mapping, with the context being designed to bias inferences in one direction or another, and the communication task allows us to uncover how participants generalise to novel meanings.

3.2 Experiment 2

3.2.1 Method

Participants

60 participants at the University of Edinburgh (39 female) were recruited via the SAGE careers database and randomly assigned to one of the possible three experimental conditions (see 3.2.1). Each condition consisted of a pair of participants who learned an artificial language and then used this language in a communication game. Participants were paid £5 for their participation.

Stimuli: Images and Input Language

Participants were asked to learn and then use an *alien language*, consisting of lower-case labels paired with images. There were 12 images that varied along three features: shape, colour, and a unique identifier (see figure 3.2). Of these 12 images, eight were randomly selected for training (*training set*) within the following constraint: the set of 8 images could not have less than 2 and more than 3 members with the same feature (e.g., figure 3.2). Each image was then assigned a label as follows: from a set of vowels (a,e,i,o,u) and consonants (g,h,k,l,m,n,p,w) we randomly generated 4 CV and 3 CVC syllables which were then assigned to 4 colours and 3 shapes. Each label therefore contained 1 initial CV syllable for a specific colour and 1 final CVC syllable for a specific shape.

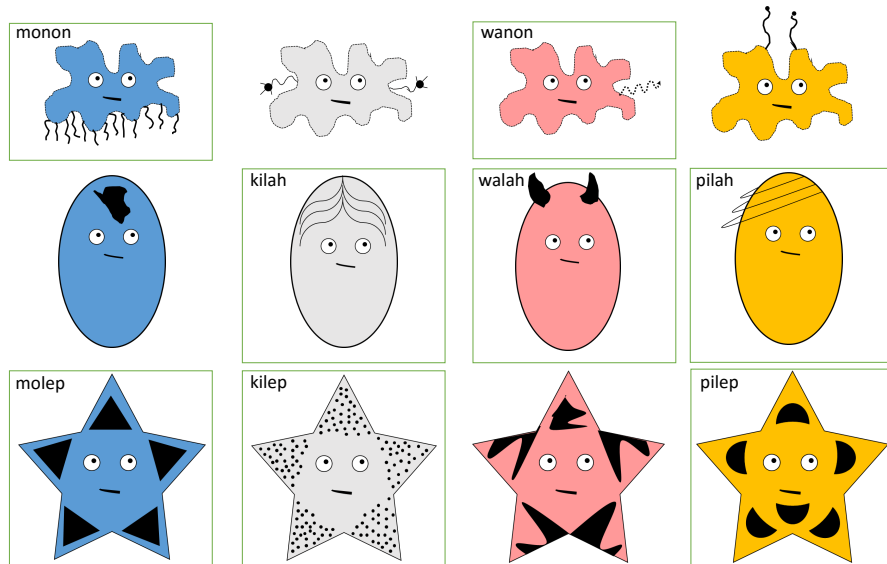


Figure 3.2: Meaning space used in Experiment 2. Images and forms inside green boxes indicate an example training set. Images without green boxes and labels were only seen during the communication phase.

Procedure

At the start of the experiment, participants were told they would first have to learn and then communicate using an alien language. Participant pairs completed the experiment in separate booths on networked computers. The experiment consisted of two main phases: a *training phase* and a *communication phase*. Before each phase began, detailed instructions were given on what that phase would involve, with a reminder not to use English or any other language during the experiment. For the training phase, participants completed the task separately, and it was only during the communication phase where they interacted (remotely, over the computer network). Both training and communication phases are the same as in Experiment 1 (see Chapter 2)

Manipulating Context-Type

To test for the role of context, a simple manipulation was made to the possible combinations of target and distractor images within a single trial during training and communication. This provides three experimental conditions based on which dimensions are relevant for successfully discriminating the target from the distractor (see Figure 3.3). For the *Shape-Different* condition, a trial consists of a target and distractor that always differ in shape, but share the same colour, whereas in the *Colour-Different* condition the reverse is true: targets and distractors are always differentiated on colour, but share the same shape. Lastly, for the

Both-Different condition, we constructed the target and distractor so that they always differed on both colour and shape.

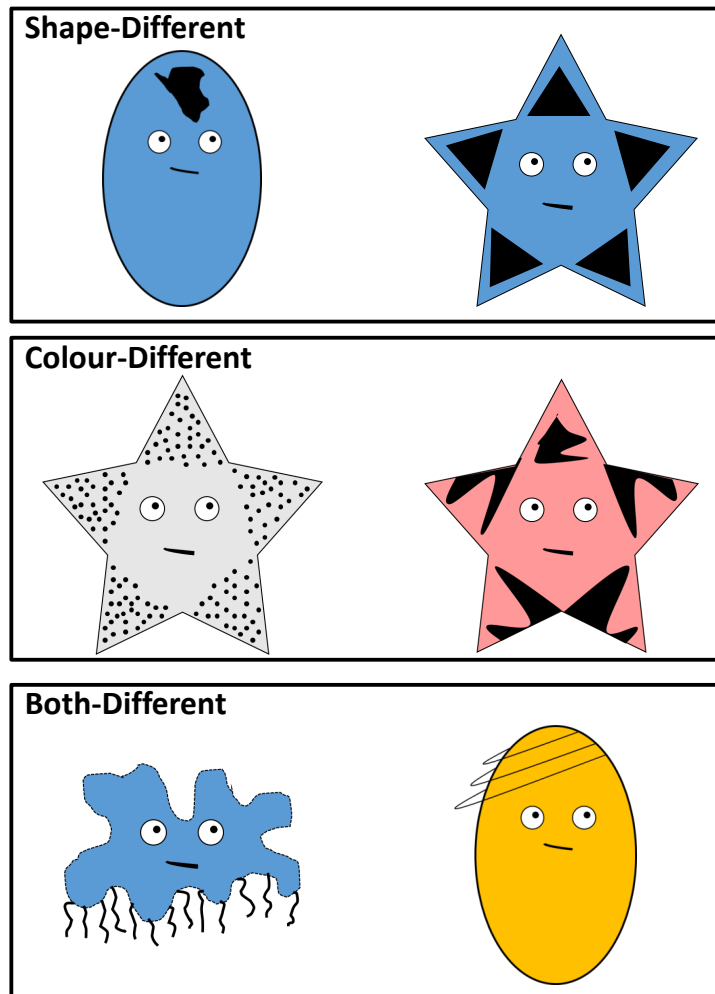


Figure 3.3: Example of the three types of referential context used in this experiment. From top to bottom: Shape-Different (Shape always differs between target and distractor, Colour is always the same), Colour-Different (Colour always differs between target and distractor, Shape is always the same), Both-Different (Shape and Colour always differ between target and distractor).

3.2.2 Dependent Variables and Hypotheses

Measuring Communicative Success

To measure communicative success we recorded the number of successful trials between the speaker and the hearer, i.e., when the hearer clicked on the target image. The maximum success score was 48 points for two blocks of 24 trials. The aim of this measure is to see whether distinct communication systems are

communicatively functional, i.e., successfully discriminating between a target and distractor above the level of chance.

Measuring Uncertainty: conditional entropy

To quantify the types of mappings between signals and meanings we measured the *conditional entropy* (Winters, Kirby & Smith, 2015; Chapter 2). However, the measures of signal uncertainty and meaning uncertainty do not tell us on what dimension participants are generalising. To resolve this problem we added a measure of feature uncertainty, $H(F|S)$, which is the expected entropy (i.e., uncertainty) over features given a signal:

$$H(F|S) = - \sum_{s \in S} P(s) \sum_{f \in F} P(f|s) \log P(f|s) \quad (3.1)$$

where the rightmost sum is the entropy over features given a particular signal $s \in S$. $P(f|s)$ is the probability that feature f is the intended feature given that signal s has been produced. This entropy is weighted by the distribution $P(s)$ on signals. F is the set of shapes and colours, and $f \subset F$ is either *Shape* = $\{f \in F | f \notin \text{Colour}\}$ or *Colour* = $\{f \in F | f \notin \text{Shape}\}$. A high $H(F|S)$ indicates there is uncertainty about a feature (within one of the two subsets) due to a signal mapping onto multiple features. It also tells us which dimension has the highest uncertainty given a signal: if a language only encodes colour, then $H(F|S)$ for the dimension of shape will be high and the $H(F|S)$ for the dimension of colour will be low (and vice versa for a language which only encodes shape).

Hypotheses

Here we provide a set of hypotheses related to our specific measurements:

Hypothesis One: Communication systems will be functionally adequate for identifying the intended meaning in context. As such, we predict all conditions will reach a communicative success score higher than chance (>50%).

Hypothesis Two: Participants in the Shape-Different condition are predicted to show greater levels of underspecification where only shape is encoded. This will result in systems with a high $H(M|S)$, a low $H(S|M)$, a high $H(F|S)$ for the Colour-Dimension and a low $H(F|S)$ for the Shape-Dimension.

Hypothesis Three: Participants in the Colour-Different condition are predicted to have greater levels of underspecification where only colour is encoded.

This will result in systems with a high $H(M|S)$, a low $H(S|M)$, a low $H(F|S)$ for the Colour-Dimension and a high $H(F|S)$ for the Shape-Dimension.

Hypothesis Four: Participants in the Both-Different condition are predicted to have higher levels of compositionality where both colour and shape are encoded. This will result in systems with a low entropy for $H(M|S)$, $H(S|M)$, and $H(F|S)$.

3.2.3 Results

Analysis: Mixed Effect Models

All analyses were performed in R (R Core Team, 2016) and used the *lme4* package (Bates, Maechler, Bolker & Walker, 2015) to run logit mixed-effect regressions for the dependent variables of communicative success and $H(S|M)$, and linear mixed-effect regressions for the dependent variables of $H(M|S)$ and $H(F|S)$. Random effects included intercepts for Participant and Initial Training Language. Each intercept had random slopes for the fixed effects and P-values for the fixed effects were obtained using the *lmerTest* package (Kuznetsova, Brockhoff & Christensen, 2014).

Communicative Success

Performance during the communication phase was extremely high in all three conditions, with the mean scores of Shape-Different ($M = 97.92\%$, 95% CI² [96.25, 99.38]), Colour-Different ($M = 84.58\%$, [76.46, 91.46]), and Both-Different ($M = 93.13\%$, [89.17, 96.25]) all being above chance (see Figure 3.4). However, as these means highlight, there appears to be an effect of condition. A logit regression with Block and Condition (reference level = Colour-Different³) as fixed effects supports this observation: communicative success is significantly higher in Shape-Different relative to Colour-Different ($\beta = 2.373$, $SE = 0.812$, $p = .004$). There is also a significant intercept ($\beta = 2.281$, $SE = 0.460$, $p < .001$) confirming performance is above chance. All other predictors and associated interactions are non significant ($p > .294$).

²All 95% CIs were bootstrapped by resampling the original data sample 10000 times using the *Hmisc package* (Harrell Jr et al., 2014). This allows us to approximate a CI for the sample statistic without making assumptions about the shape of the distribution (it is nonparametric).

³There was no a-priori hypothesis specified about differences in communicative success based on condition. For exploratory purposes, we used Colour-Different as the reference level, given the unexpected observation that its mean was the lowest.

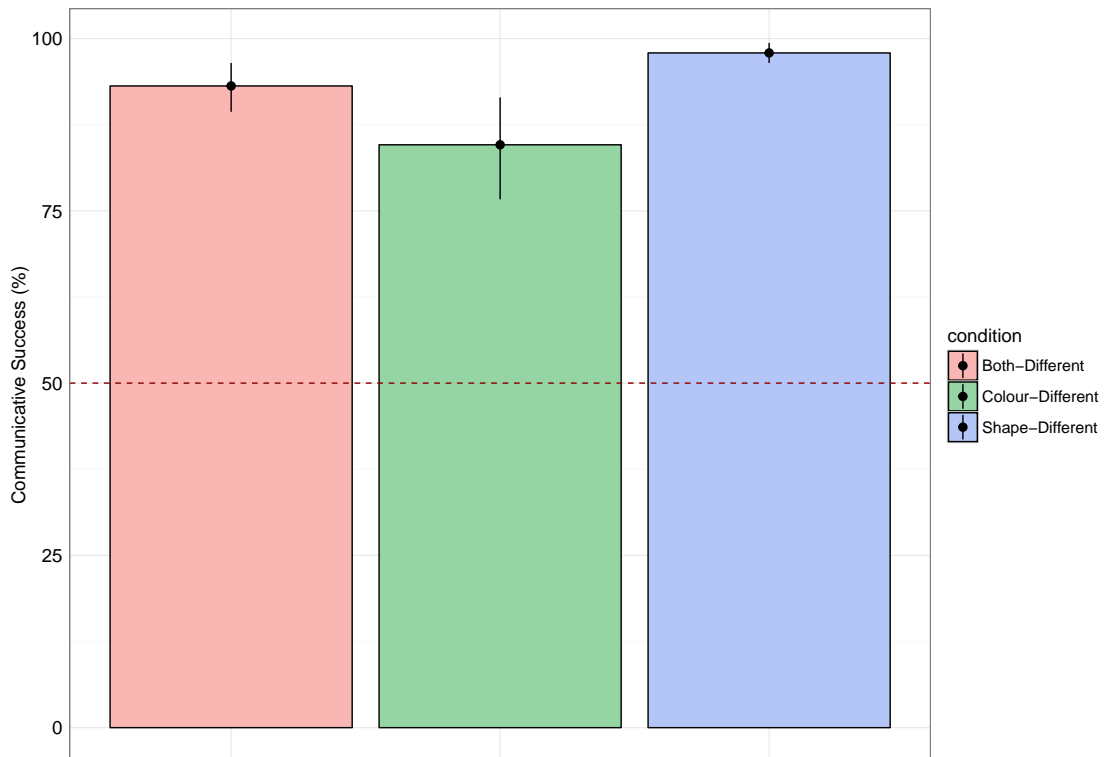


Figure 3.4: Average communicative success scores by condition. Error bars represent bootstrapped 95% Confidence Intervals. Red dotted line is the chance level of communicative success (50%).

Signal Uncertainty

Figure 3.5 shows the degree of signal uncertainty, measured by $H(S|M)$, for the three experimental conditions. As the visual inspection suggests, participants in Colour-Different ($M = 0.27$, $[0.21, 0.33]$) produced considerably less variable signals when compared with those in Shape-Different ($M = 0.56$, $[0.5, 0.63]$) and Both-Different ($M = 0.49$, $[0.44, 0.56]$). This is unexpected given our prediction that all three conditions should have low signal uncertainty. A logit regression with Block and Condition (reference level = Both-Different) indicates that participants in Colour-Different ($\beta = -1.343$, $SE = 0.589$, $p = .023$) have significantly lower levels of signal uncertainty. All other predictors and associated interactions are non-significant ($p > .598$).

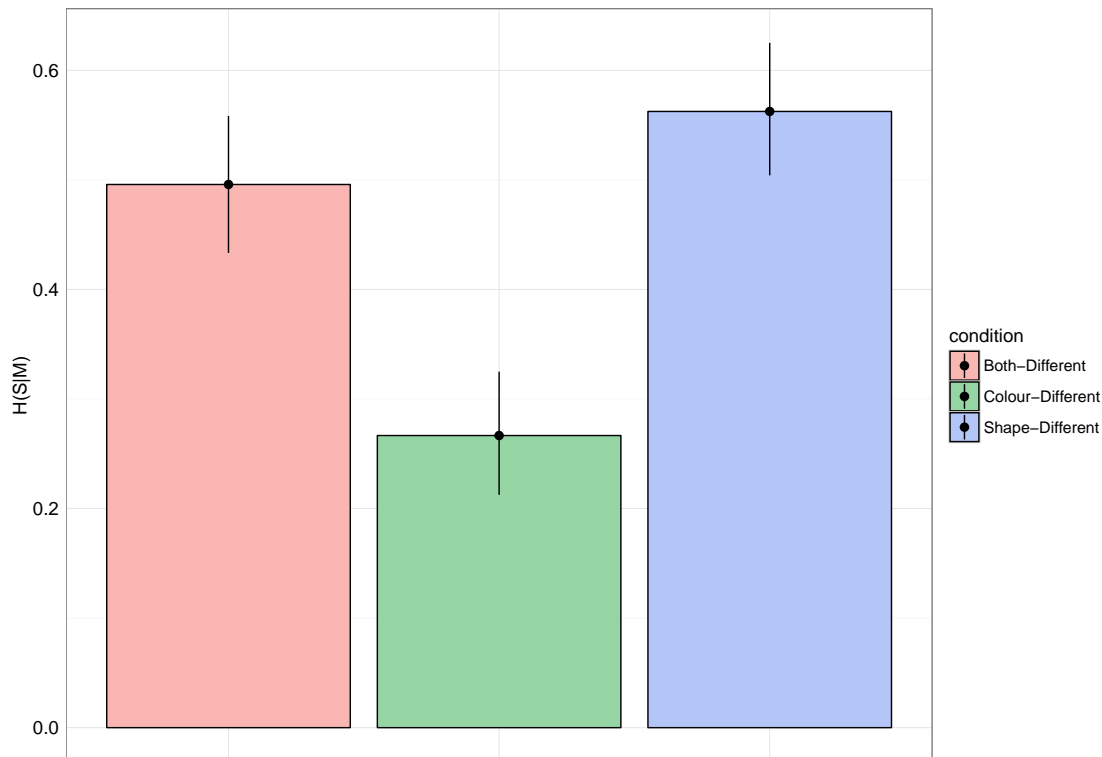


Figure 3.5: Degree of signal uncertainty, measured by $H(S|M)$. Higher entropy values indicate a higher degree of signal uncertainty. The errors represent bootstrapped 95% CIs.

Meaning Uncertainty

Figure 3.6 shows the extent to which signals are used across multiple meanings, measured by $H(M|S)$. Our hypothesis is that both the Shape-Different and Colour-Different conditions should show higher levels of meaning uncertainty than the Both-Different condition. This is somewhat confirmed by the higher average for Shape-Different ($M = 0.42$, $[0.35, 0.51]$) than Both-Different ($M = 0.34$, $[0.28, 0.41]$). One unexpected result is that the average meaning uncertainty for Colour-Different is the lowest ($M = 0.12$, $[0.08, 0.18]$) and suggests participants in this condition are underspecifying less than predicted. A linear mixed effect model with Block and Condition (reference level = Both-Different) as fixed effects shows that Shape-Different ($\beta = 0.140$, $SE = 0.125$, $p = .004$) is a significant predictor of $H(M|S)$. What this tells us is that participants in Shape-Different are more likely to reuse signals across multiple meanings when compared with participants in Both-Different. However, there is only a marginal difference between Both-Different and Colour-Different conditions ($\beta = -0.253$, $SE = 0.127$, $p = .054$). All other predictors and associated interactions are non-significant ($p > .182$).

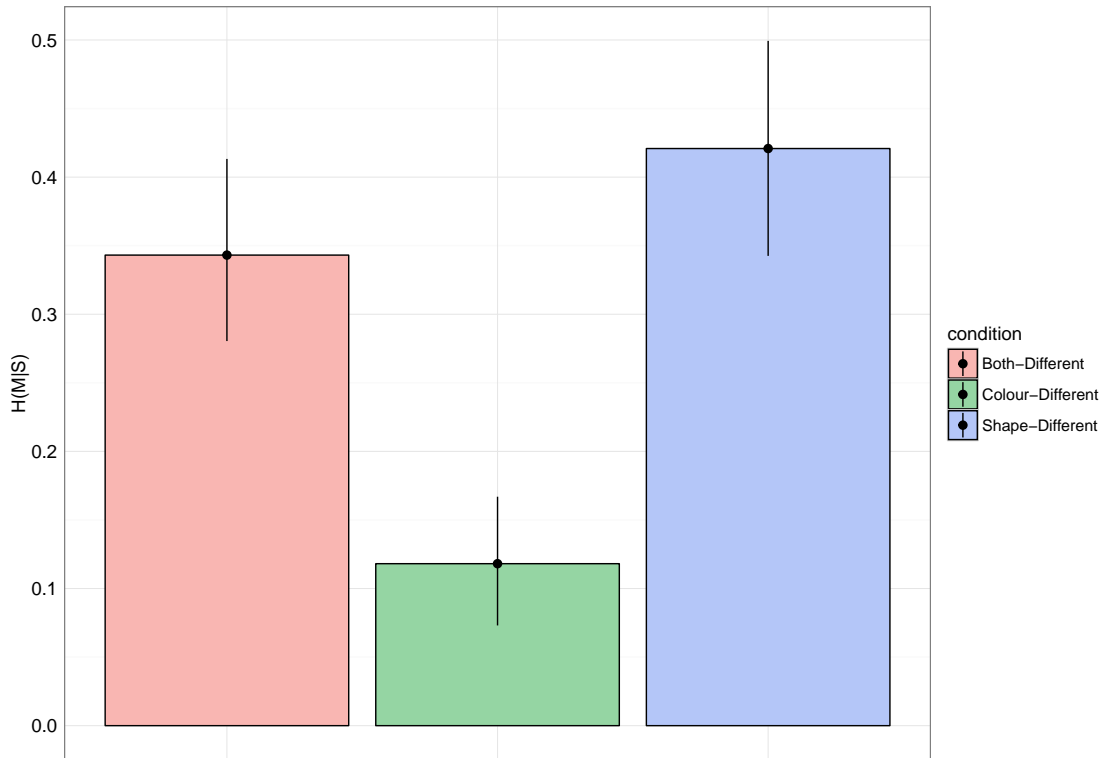


Figure 3.6: Degree of meaning uncertainty, measured as $H(M|S)$.

Feature Uncertainty

There are two interesting aspects about the feature uncertainty, measured by $H(F|S)$ (see Figure 3.7). The first is that there is generally lower feature uncertainty for the Shape-Dimension ($M = 0.03$, $[0.02, 0.04]$) than the Colour-Dimension ($M = 0.28$, $[0.24, 0.31]$). This suggests a general preference for encoding shape over colour. The second point is that this preference tends to be modulated by condition; indicated by the clear difference in means for Shape-Different (Shape-Dimension: $M = 0.02$, $[0.00, 0.03]$; Colour-Dimension: $M = 0.43$, $[0.35, 0.51]$) and Both-Different (Shape-Dimension: $M = 0.04$, $[0.02, 0.06]$; Colour-Dimension: $M = 0.32$, $[0.26, 0.39]$) when compared with Colour-Different (Shape-Dimension: $M = 0.04$, $[0.02, 0.07]$; Colour-Dimension: $M = 0.09$, $[0.06, 0.14]$).

A linear mixed-effect model, with Block, Condition (reference level = Both-Different) and Dimension (reference level = Colour-Dimension) as fixed effects, confirms that both Shape-Dimension ($\beta = -0.276$, $SE = 0.046$, $p < .001$) and Colour-Different condition ($\beta = -0.13$, $SE = 0.062$, $p = .039$) are significant predictors of $H(F|S)$. The significant interactions for Colour-Different x Shape-Dimension ($\beta = 0.222$, $SE = 0.064$, $p < .001$) and Shape-Different x Shape-Dimension ($\beta = -0.151$, $SE = 0.066$, $p = .022$) also provides confirma-

tion that the shape bias is modulated by condition. First, the interaction for Colour-Different x Shape-Dimension tells us that there is little difference in feature uncertainty for Shape- and Colour-Dimensions in Colour-Different (when compared to the difference between these dimensions for Both-Different). The fact that both Colour and Shape have low feature uncertainty suggests that participants in Colour-Different are not underspecifying. Second, the interaction for Shape-Different x Shape-Dimension shows that the difference in feature uncertainty for Shape- and Colour-Dimensions is higher than that in Both-Different: that is, participants in Shape-Different are less likely to encode colour than participants in Both-Different. All other predictors and associated interactions are non-significant ($p > .278$).

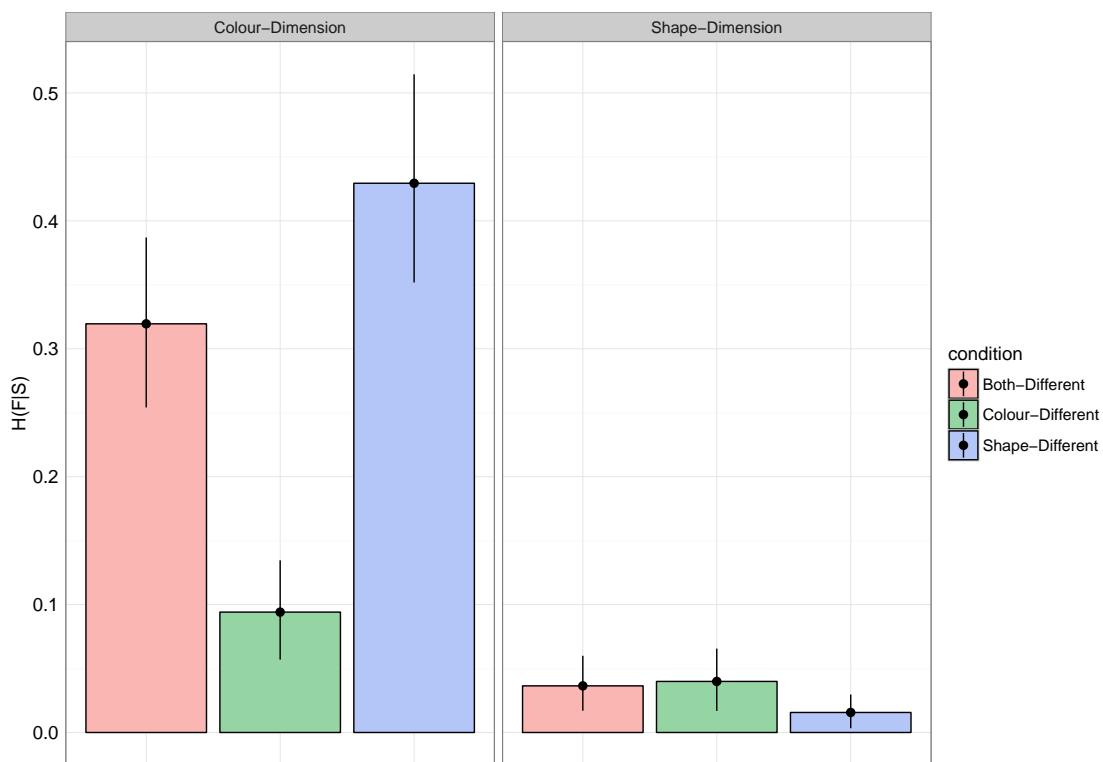


Figure 3.7: Degree of feature uncertainty, measured as $H(F|S)$, for the dimensions of Shape and Colour.

3.2.4 Experiment 2 Discussion

Experiment 2 provides tentative support for the hypothesis that contextual constraints can disrupt an initially compositional system. The results show that (i) the systems used by participants are communicatively functional in discriminating between meanings (as indicated by high communicative accuracy across condi-

tions); (ii) participants in Shape-Different and Both-Different conditions tend to produce more variable signals (i.e., higher signal uncertainty) and show higher levels of signal reuse across multiple meanings (i.e., higher meaning uncertainty); (iii) signal reuse in Shape-Different and Both-Different conditions tends to underspecify on colour and encode shape (i.e., high feature uncertainty for the colour dimension).

The key finding is the unexpected interaction between our manipulations to the referential context and a bias to encode shape. This results in more underspecification when shape is a relevant dimension for discriminating between meanings. Furthermore, this effect is amplified when shape is the only relevant dimension and colour is backgrounded: that is, participants in Shape-Different tend to underspecify more than participants in Both-Different. Figure 3.8 shows an example of an underspecified language from the Shape-Different condition: if participants are using the compositional rule, then we would expect them to produce the forms *nimel*, *hawuh*, *pukup*, and *kakup* for the unseen set, where in fact this pair reused forms from their training data (e.g., *nikup*, which originally meant grey (*ni*) star (*kup*)) to convey shape (e.g., *nikup* is now used for grey, pink, orange and blue stars). This mirrors what linguists observe in the historical record: an initially compositional construction is chunked, reanalysed as a single linguistic unit and then generalised to convey novel meanings.

By contrast, participants in Colour-Different overwhelmingly preferred to maintain compositional systems (see Figure 3.9). This suggests that our prior assumptions are wrong, with the results for Colour-Different most likely being the product of two competing pressures: a contextual bias (to discriminate on the basis of colour) and a shape bias (to use shape as a categorisation cue). Cases where we do see underspecification in Colour-Different also tend to encode shape, and partly explains why participants in this condition have lower average success scores.

What do these results mean for our predictions? It ultimately depends on the source of this preference to encode shape over colour. One explanation is that there is a shape bias which interacts with our manipulations to the referential context. For instance, when extending object names, children tend to do so on the similarity of shape as opposed to other features, such as size, colour or material (Smith et al., 2002; Diesendruck & Bloom, 2003; Elman, 2008). Still, the claim as to the nature of a shape bias, let alone the specific details concerning the strength of this bias and its underlying mechanisms, is disputed (for review, see: Elman, 2008). Alternatively, the shape bias could be attributable to another source, such as the native language of the participants: that is, participants are mapping their

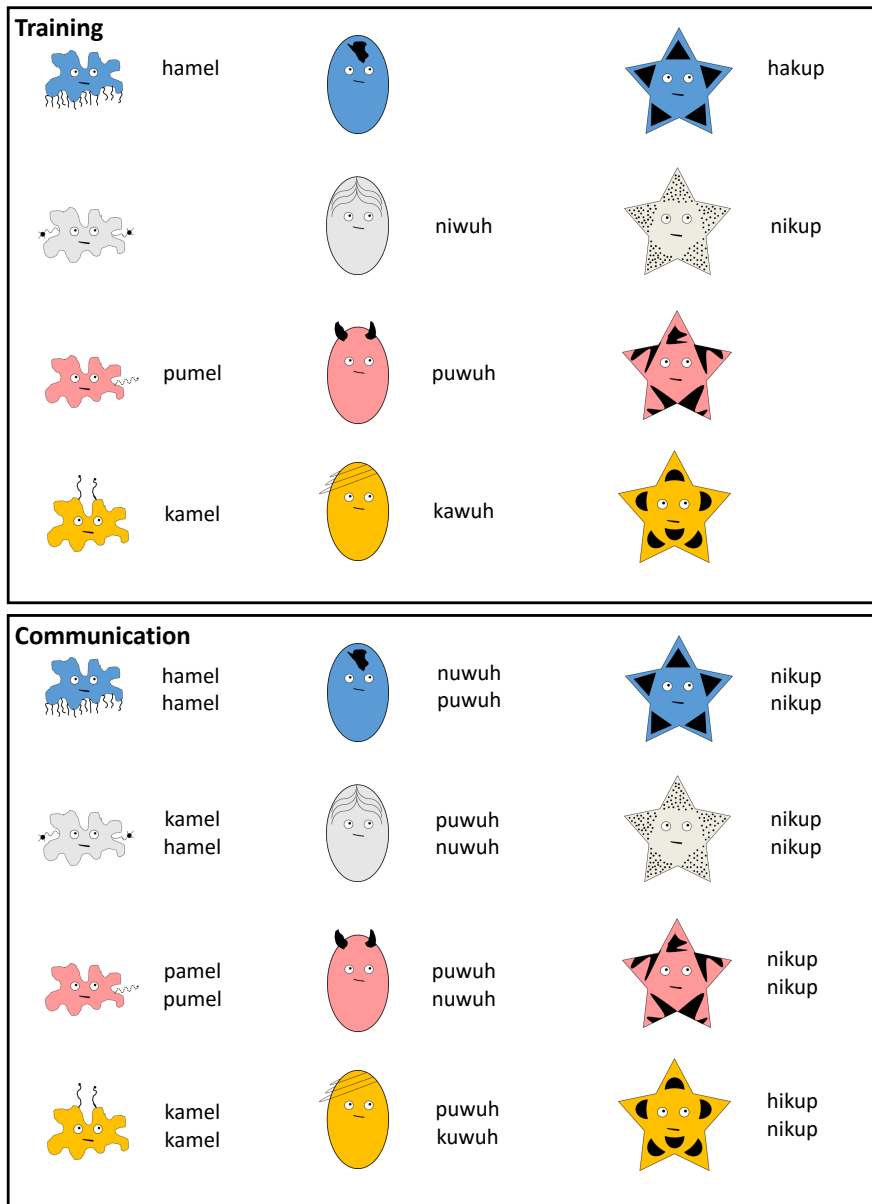


Figure 3.8: An example of an Underspecified language in the Shape-Different condition. The top box shows the signals and meanings participants were trained on and the bottom box shows the forms participants produced during communication.

rules for English onto what is and is not permissible in the artificial language. For English speakers, dropping the adjective in the sentence *The red square* is perfectly valid in terms of grammaticality (i.e., *The square*), whereas dropping the noun is considered ungrammatical (i.e., **The red*). It could be the case that this alters the participants' expectation about what should and should not be encoded.

If the shape bias is explained by either a perceptual or native-language bias,

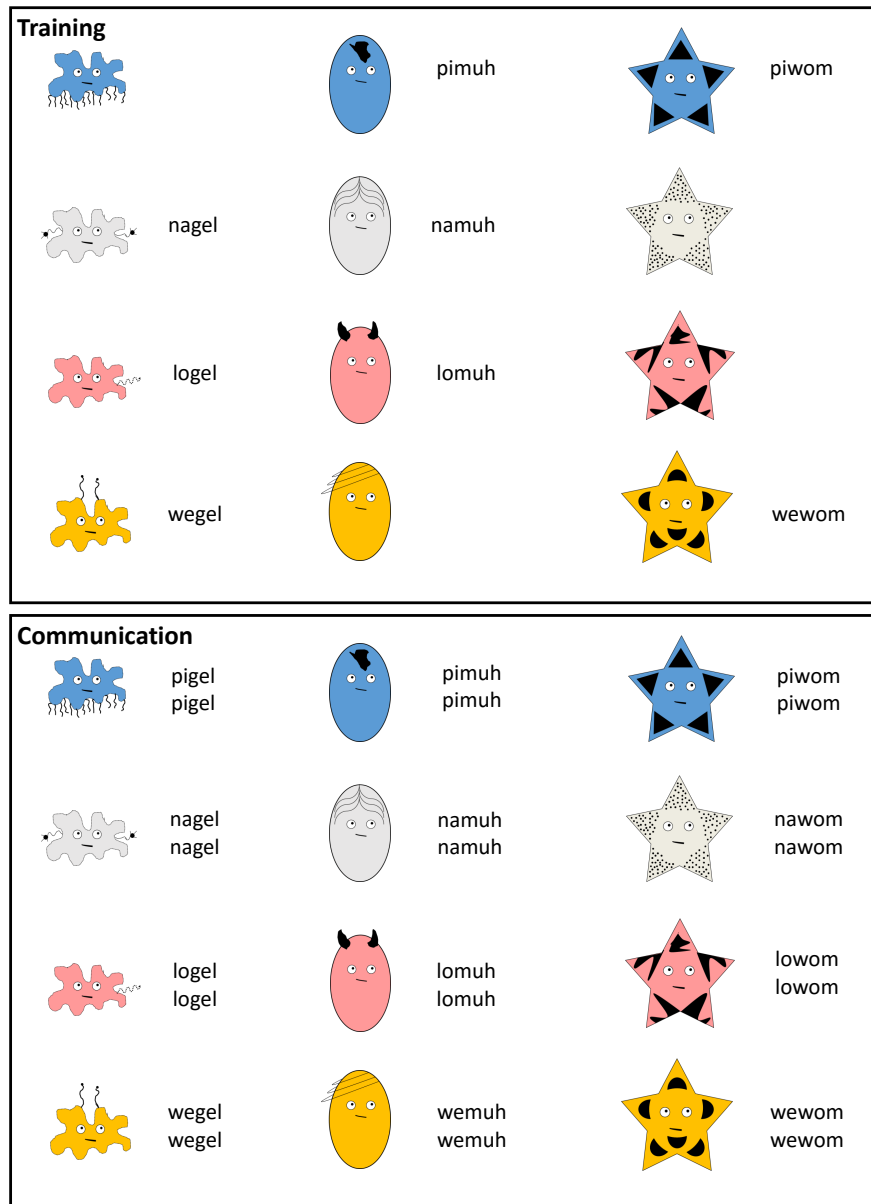


Figure 3.9: An example of a compositional language in the Colour-Different condition. The top box shows the signals and meanings participants were trained on and the bottom box shows the forms participants produced during communication.

then this nicely explains our experimental results: both the shape bias and contextual bias interact with one another, influencing the probability of a participant learning and using a compositional mapping in discrimination. In the Shape-Different condition, the prior shape bias and the contextual bias are reinforcing, which results in a strong pressure to generalise and only encode shape. By contrast, the Colour-Different condition has a potential conflict between the shape bias (which favours the encoding of shape) and the discriminatory need to convey

colour (as this is the only dimension relevant for discrimination). It is possible these conflicting biases decrease uncertainty in interpretation by highlighting both dimensions, making it more probable that participants will learn and maintain the compositional system. Lastly, for the Both-Different condition, the contextual bias to convey shape might not be as strong, as colour is also a relevant feature, but if participants do underspecify then the shape bias provides an added incentive to preference shape over colour (and this is what we see on the basis of feature uncertainty).

The last possibility is a confound in the experimental design: due to there being fewer shapes (Blob, Oval, Star) than colours (Blue, Grey, Pink, Yellow), a bias was introduced to encode the dimension with the fewest features. This still leaves open the following question: how does having fewer features relate to underspecification? One possibility is that it is easier to learn a mapping where the set of signals map onto 3 shapes than it is to learn a mapping where the set of signals map onto 4 colours. Under this account, there is nothing intrinsically special about shape, with participants simply being more likely to underspecify on the dimension with the fewest features. There is already evidence to suggest that, when the dimensions are balanced, the preference for one dimension over another is contingent on history (Kirby, Cornish & Smith, 2008) and context (Silvey, Kirby & Smith, 2015). Under this account, we should expect that switching the skew in the opposite direction, so that there are 4 shapes and 3 colours, will create a bias to convey the colour dimension. Furthermore, this skew also introduces a frequency bias in training: participants are provided with more evidence about how signals map onto shapes than colours. These two aspects potentially explain why participants are more likely to underspecify when shape is relevant for discrimination: having fewer shapes in the set, as well as having a greater exposure to shapes across trials, supports the hypothesis that signals only map onto the shape dimension. Similarly, the reason why participants maintain compositional languages in Colour-Different is because there is a functional pressure to learn colour (for discrimination) and a skewness bias to learn shape (as there are fewer shapes). That is, in terms of learning, shape is a useful categorisation cue even when it is not useful for discrimination.

We explicitly test these competing hypotheses in Experiment 3 by counterbalancing for the skew in the number of dimensions. The general methodology is the same as in this experiment except each condition now has two subsets: one group of participants is exposed to a meaning space of 4 colours and 3 shapes (Colour-Skewed) and the other group is exposed to a meaning space of 3 colours and 4 shapes (Shape-Skewed). In particular, we were interested in whether there

is a bias to encode the dimension with fewer features, and to see if this interacts with a contextual bias for discriminating between meanings.

3.3 Experiment 3

3.3.1 Shape Bias or Unbalanced Dimensions?

In Experiment 2 we found that, when communication systems did transition from a compositional to an underspecified one, there was a preference to specify the shape dimension over the colour dimension. Importantly, this bias to specify shape was not restricted to the Shape-Different condition, as our original hypotheses had predicted; it was also detected in Both-Different and, to a much lesser extent, Colour-Different conditions. Based on our experimental set up, there are two possible sources for this bias: either shape is simply a better categorisation cue than colour (*Shape Bias hypothesis*) or the preference for shape is an artefact of the skew in the number of features (*Skewness Hypothesis*).

To test these two hypotheses we counterbalanced across all three conditions for the number of features: half of the participant pairs are presented with meaning spaces with a *Colour Skew* (4 colours, 3 shapes) and the other half are presented with meaning spaces with a *Shape Skew* (3 colours, 4 shapes) (see 3.3.2 for further details). A Shape Bias Hypothesis predicts this manipulation to have little effect, with the results mirroring those in Experiment 2: underspecification will only take place when shape is the relevant dimension for discrimination. By contrast, if the bias is due to a skew in the number of features (Skewness Hypothesis), then the expectation is that participants will only underspecify when the dimension with the fewest features is also relevant for discrimination. Lastly, there is the possibility of an interaction between the Shape-Bias and Skewness, which predicts underspecification only when shape is relevant for discrimination and is the dimension with the fewest features (Colour-Skewed).

3.3.2 Method

Participants

96 participants at the University of Edinburgh (67 female) were recruited via the SAGE careers database and randomly assigned to one of the possible three experimental conditions (see 3.2.1). Each condition consisted of a pair of participants who learned an artificial language and then used this language in a communication game.

Stimuli

There were 16 images that varied along three features: shape, colour, and a unique identifier (see Figure 3.10). One feature was removed depending on whether participants were in Shape-Skewed or Colour-Skewed (see 3.3.2 for more details). Image selection and generation of labels was the same as in Experiment 2 (see section 3.2.1). Note that a possible confound was unintentionally introduced with these new stimuli in that they are brighter in colour than those in Experiment 2 (see 3.3.4).

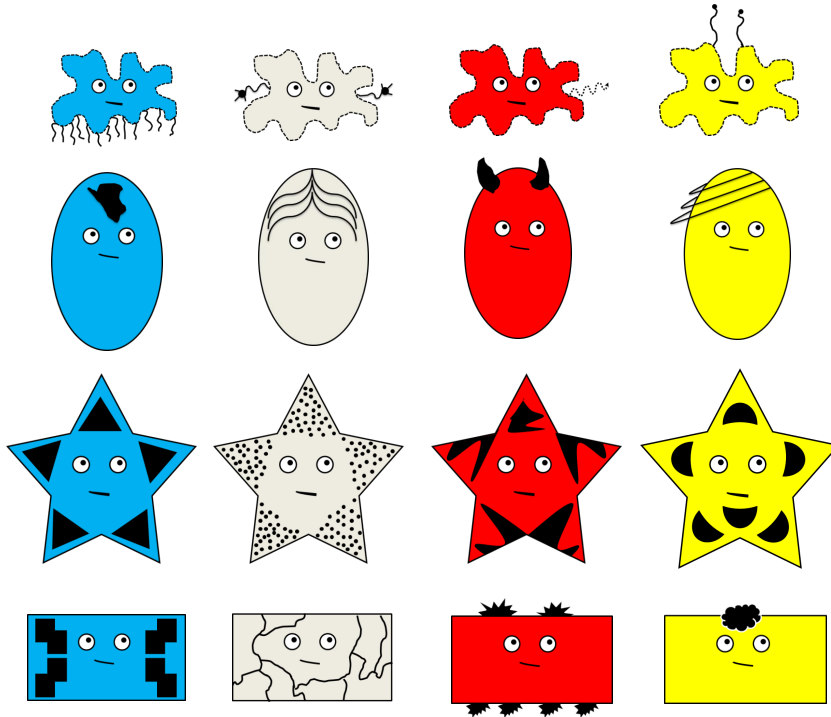


Figure 3.10: The full meaning space used in this experiment. Note that in the experiment one feature was removed depending on whether participants were in Shape-Skewed (i.e., one colour would be removed, e.g., yellow) or Colour-Skewed (i.e., one shape would be removed, e.g., star).

Procedure

The procedure is identical to Experiment 2 (see 3.2.1).

Controlling for Unbalanced Dimensions

To control for the unbalanced dimensions we introduced a new shape and manipulated the skew in the number of dimensions. This resulted in two possible

meaning spaces based on the number of shapes and colours. If the counterbalance was *colour-skewed*, then there were 3 shapes and 4 colours, whereas for the *shape-skewed* counterbalance there were 4 shapes and 3 colours. This manipulation was done systematically across all three conditions: within each condition 8 participant pairs were exposed to a colour-skewed meaning space and 8 participant pairs were exposed to a shape-skewed meaning space. All possible meaning spaces were covered within the constraints of the skew.

3.3.3 Results

Communicative Success

Performance during the communication phase was extremely high in all three conditions, with the mean scores of Shape-Different ($M = 94.01\%$, 95% CI [89.32, 97.79]), Colour-Different ($M = 93.10\%$, [89.19, 96.48]), and Both-Different ($M = 92.71\%$, [89.32, 95.83]) all above chance (see Figure 3.11). A Logit regression with Block, Condition (reference level = Both-Different) and Counterbalance (reference level = Colour-Skewed) as fixed effects confirms performance is above chance ($\beta = 3.412$, $SE = 0.696$, $p < .001$) and that there are no significant effects for any of these predictors ($p > .127$). Like the second experiment, it seems all three conditions have communicatively functional systems, in that a signal is capable of identifying the intended meaning in context. Although, in contrast to the second experiment, all of the conditions appear to reach similar levels of communicative success.

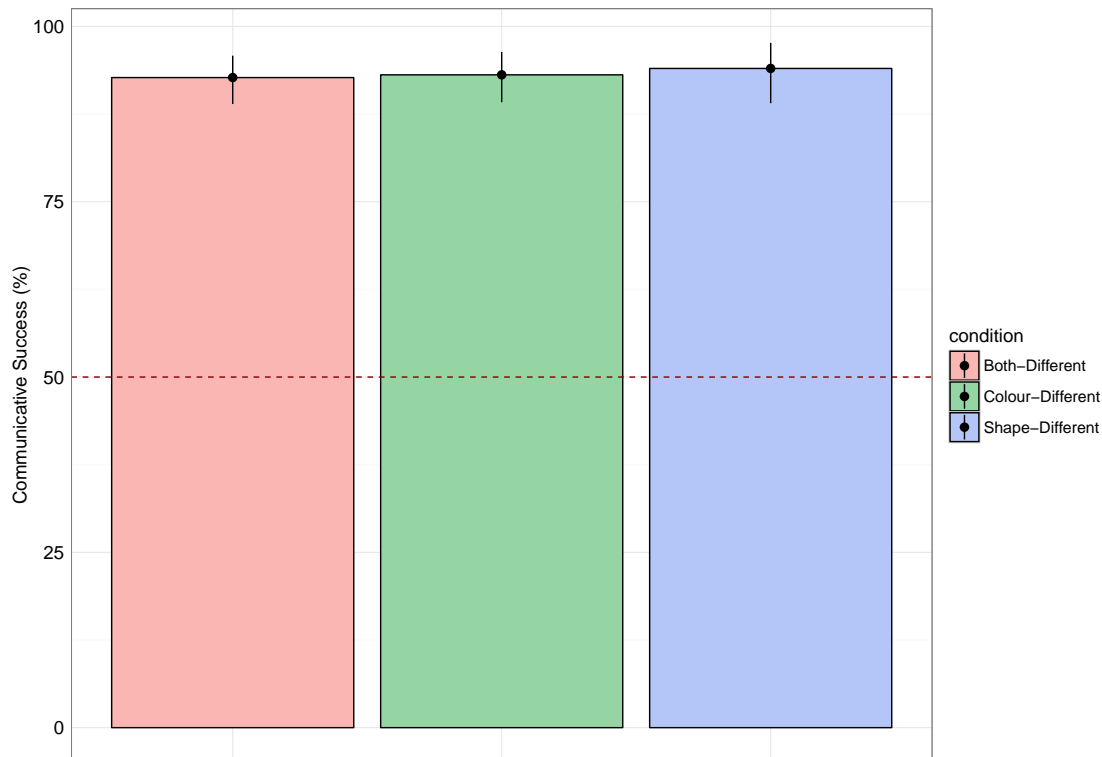


Figure 3.11: Average communicative success scores by condition. Error bars represent bootstrapped 95% Confidence Intervals for 10000 runs. Red dotted line is the chance level of communicative success (50%).

Signal Uncertainty

Signal uncertainty in Experiment 3 shows that participants in the Shape-Different condition ($M = 0.32$, $[0.28, 0.36]$) appear to produce lower levels of signal uncertainty than participants in the Both-Different ($M = 0.49$, $[0.44, 0.54]$) and Colour-Different ($M = 0.45$, $[0.41, 0.51]$) conditions (Figure 3.12). In a model with Condition (reference level = Both-Different), Block and Counter-Balance (reference level = Colour-Skewed) as fixed effects, both Colour-Different ($\beta = -0.715$, $SE = 1.043$, $p = .493$) and Shape-Different ($\beta = -0.477$, $SE = 1.039$, $p = .646$) are non-significant predictors of $H(S|M)$ when compared to Both-Different. This suggests signal variability is not dependent on condition (when controlling for participants and initial language). All other predictors and associated interactions are non-significant ($p > .147$).

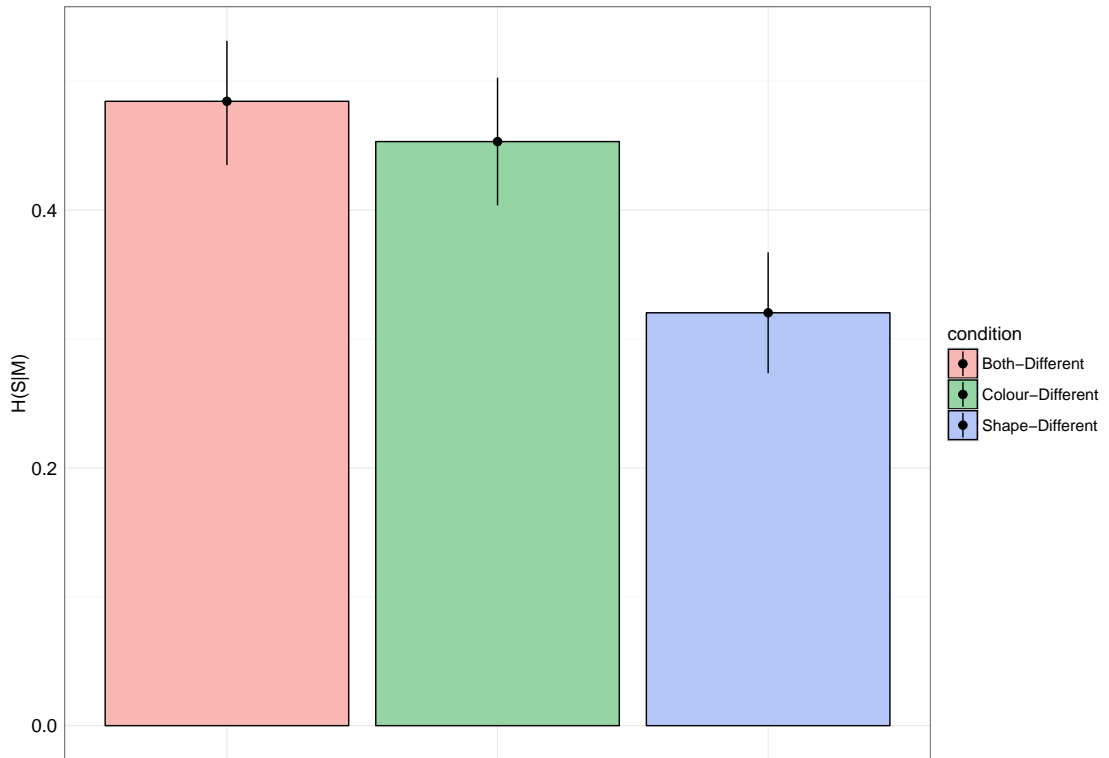


Figure 3.12: Degree of signal uncertainty, measured by $H(S|M)$. Higher entropy values indicate a higher degree of signal uncertainty. The errors represent bootstrapped 95% CIs.

Meaning Uncertainty

Figure 3.13 shows the degree of meaning uncertainty produced by participants. The first point to note is that the average meaning uncertainty in Experiment 3 ($M = 0.17$, $[0.15, 0.20]$) is lower than in Experiment 2 ($M = 0.29$, $[0.25, 0.33]$). In short, participants are generalising less across meanings, irrespective of condition. Furthermore, the trend appears to go in the opposite direction to Experiment 2, with Shape-Different ($M = 0.10$, $[0.07, 0.14]$) now having lower levels of meaning uncertainty when compared to Colour-Different ($M = 0.23$, $[0.18, 0.28]$) and Both-Different ($M = 0.18$, $[0.15, 0.23]$). There is a significant interaction between Colour-Different x Block x Shape-Skewed ($\beta = 0.320$, $SE = 0.110$, $p = .004$) in a model with Condition (reference level = Both-Different), Block and Counter-Balance (reference level = Colour-Skewed) as fixed effects. When there are more shape features (Shape-Skewed), participants in Colour-Different (when compared with Both-Different) tend to increase the degree of meaning uncertainty from the first to the second block. This simply tells us that participants in Colour-Different are more likely to underspecify (i.e., reusing the same label across multiple mean-

ings) when there are more shapes than colours – and this increasingly happens in the second block. All other predictors and associated interactions are non-significant ($p > .219$).

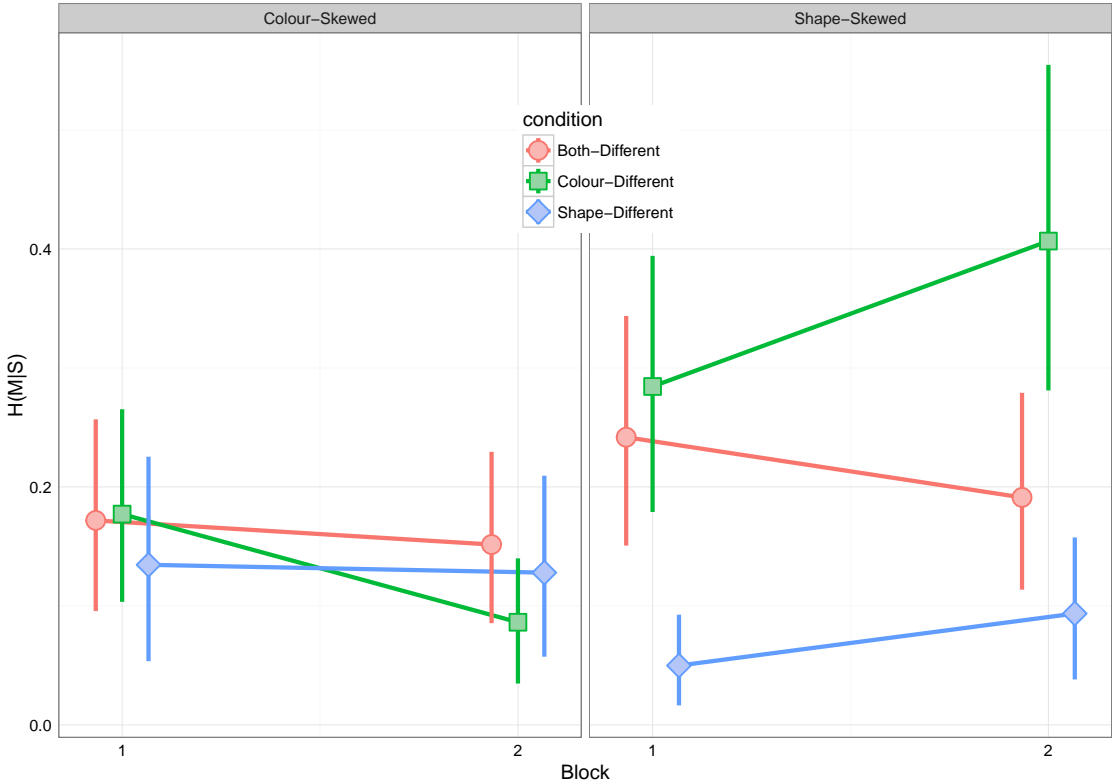


Figure 3.13: Degree of meaning uncertainty, measured as $H(M|S)$, for participant pairs in Colour-Skewed (4 Colours, 3 Shapes) and Shape-Skewed (3 Colours, 4 Shapes).

Feature Uncertainty

Figure 3.14 shows the degree of feature uncertainty for both the Colour and Shape dimensions. Inspection of the averages reveals that Colour-Different has higher feature uncertainty for Shape ($M = 0.20$, $[0.16, 0.25]$) than Colour ($M = 0.07$, $[0.04, 0.09]$); participants are therefore more likely to underspecify and encode colour. By contrast, there tends to be little difference in feature uncertainty for Both-Different (Colour Dimension: $M = 0.13$, $[0.10, 0.17]$; Shape Dimension: $M = 0.10$, $[0.07, 0.13]$) and Shape-Different (Colour Dimension: $M = 0.07$, $[0.05, 0.10]$; Shape Dimension: $M = 0.04$, $[0.02, 0.06]$). A model with Condition (reference level = Both-Different), Block, Dimension (reference level = Colour Dimension) and Counter-Balance (reference level = Colour-Skewed) as fixed effects shows there are significant interactions for Colour-Different x Shape Dimension

x Shape-Skewed x Block ($\beta = 0.285$, $SE = 0.119$, $p = .017$). As such, when there are more shapes (Shape-Skewed) in the Colour-Different condition, there tends to be higher feature uncertainty for the Shape Dimension, with this uncertainty increasing from the first to the second block. That is, when there are more shape dimensions than colour dimensions, participants in the Colour-Different condition are less likely to encode shape in their linguistic system. Furthermore, participants are more likely to not convey shape as they transition from the first to the second block.

By contrast, the significant interaction for Shape-Different x Shape Dimension x Shape-Skewed ($\beta = -0.179$, $SE = 0.083$, $p = .031$) indicates that participants in the Shape-Different condition tend to produce lower levels of feature uncertainty when there are more shape features. Participants in the Shape-Different condition therefore tend to maintain the encoding of shape (and colour) in their linguistic systems. Shape Dimension ($\beta = -0.132$, $SE = 0.041$, $p = .001$) is also a significant predictor of $H(F|M)$ and there are significant interactions for Colour-Different x Shape Dimension ($\beta = 0.227$, $SE = 0.059$, $p < .001$), Shape Dimension x Shape-Skewed ($\beta = 0.229$, $SE = 0.056$, $p < .001$), and Colour-Different x Shape Dimension x Shape-Skewed ($\beta = -0.171$, $SE = 0.084$, $p = .041$). All other predictors and associated interactions are non-significant ($p > .093$).

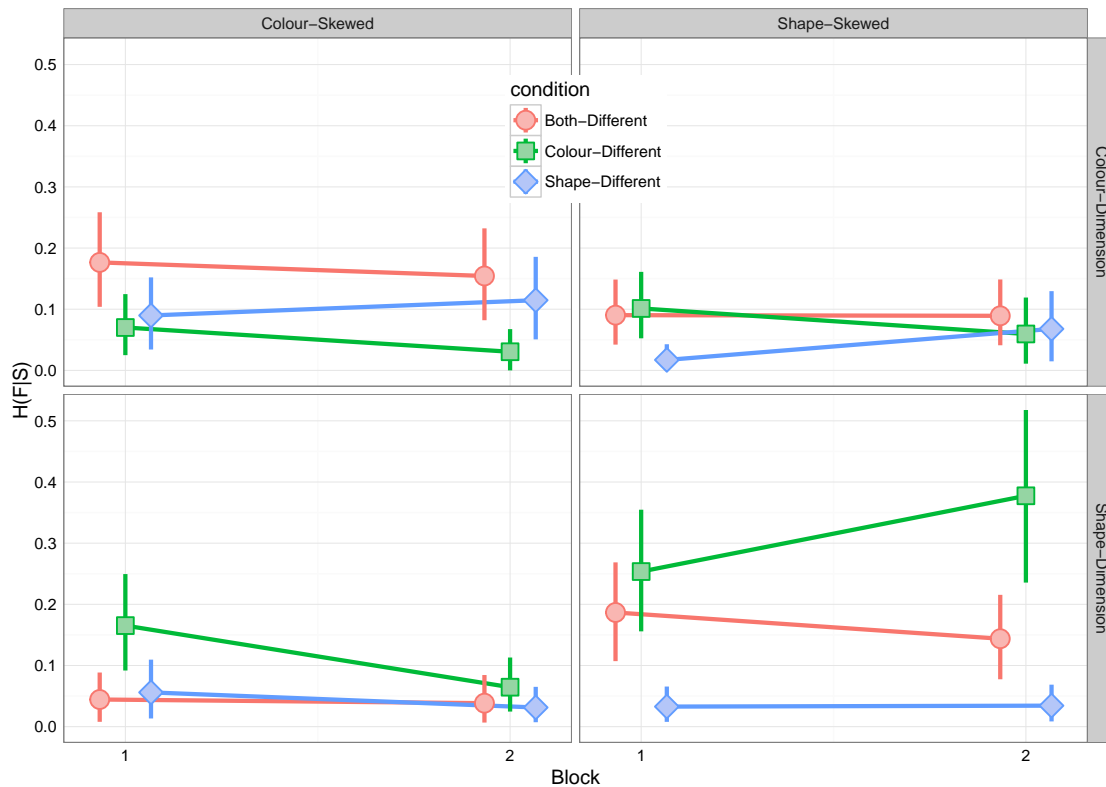


Figure 3.14: Degree of feature uncertainty, measured as $H(F|S)$, for the dimensions of Shape and Colour.

3.3.4 Experiment 3 Discussion

The results of Experiment 2 and Experiment 3 suggest that considerably more work is needed to test the general hypothesis that context motivate the loss of compositionality. The only reproducible result is communicative success: hearers successfully clicked on the target image at levels higher than chance. Signal uncertainty, meaning uncertainty, and feature uncertainty all differed in fundamental ways from the first experiment. The two most notable results being (i) the drop in overall effect size for meaning uncertainty and feature uncertainty, and (ii) the disappearance of an effect for shape. Taken together, these two results suggest the effect of context is relatively small, with the results from Experiment 2 mostly likely being an artefact of the experimental design (see 3.4 for further discussion).

The drop in effect size for meaning uncertainty and feature uncertainty tell us that participants across all conditions are underspecifying less: when compared with the original experiment, participants are more likely to learn and generalise the compositional rule, irrespective of experimental condition. Not only did the effect size drop, the direction of effect went in the opposite direction: participants

tend to underspecify more in Colour-Different than in Shape-Different, resulting in systems that only encoded colour. Even with this change in both the size and direction of the effect, context is still a significant predictor: when systems underspecify they will overwhelmingly convey the dimension relevant for discriminating between the target and distractor (i.e., if colour is the relevant dimension, and a pair underspecifies, then only colour will be encoded in the linguistic system). It appears that manipulations to context do partly motivate the loss of compositionality – albeit through interacting with other biases.

This brings us to the rationale for counterbalancing: What do these results mean for our two competing hypotheses of *shape bias* and *skewness*? Given the general trend not to encode shape in Experiment 3 it seems highly unlikely that the shape bias was driving the effect in Experiment 2. In fact, the results almost go in exactly the opposite direction to the Shape Bias predictions: Shape-Different tended to maintain compositional systems, Colour-Different had the highest proportion of underspecified systems and these encoded colour, and of the underspecified systems in Both-Different some only encoded colour and some only encoded shape. As such, the bias in Experiment 2 was most likely due to the experimental design, where the meaning space was always Colour-Skewed, and not due to an intrinsic preference for shape as a categorisation cue.

The skewness hypothesis (SH) performs somewhat better in that it does explain the results for Colour-Different and Both-Different. First, SH predicts a bias to underspecify and encode the dimension with the fewest features. For example, participants in the Both-Different condition tend to encode shape (and ignore colour) when there are more colours, and encode colour (and ignore shape) when there are more shapes. The second prediction of SH is that this bias should interact with the context by either amplifying or dampening the effect. This is what we find in the Colour-Different condition: participants disproportionately underspecify and encode colour when there is shape-skew (reinforcing bias), whereas when there is a colour-skew they tend to maintain a compositional system (conflicting bias).

The notable omission is the Shape-Different condition: In this case, participants did not fully follow what the SH would predict, with there being a general tendency to maintain compositional systems – irrespective of whether or not the bias was reinforcing or conflicting. One possibility for this outcome is a bias to encode colour that is independent of the skewness manipulation. The source of this bias could be the result of an experimental confound: in the third experiment the colours of the stimuli are brighter than those in the second experiment. Brightness is known to have a visual saliency effect (Milosavljevic et al., 2012) and

this might explain the preference to maintain compositional systems. An effect of brightness would also explain why Colour-Different tended to have higher levels of underspecification. However, it does not explain the pattern of underspecified systems in Both-Different: the expectation is that there would be a greater number of systems that only encoded colour; instead, there are a roughly equal number of systems that underspecify on colour and shape, with these patterns being predicted by skewness (although we are dealing with a relatively small effect here, so this interpretation should be treated with caution).

Still, neither skewness nor the shape bias hypotheses predicted the overall drop in effect size from Experiment 2 to Experiment 3 (i.e., participants are underspecifying less in Experiment 3 than in Experiment 2). Future work will need to investigate whether this is due to an experimental confound (e.g., colour brightness) or some other experimental manipulation. A parsimonious explanation is that both experimental results are partly due to low statistical power (for a recent discussion on these issues, see: Open Science Collaboration, 2015): i.e., there were not enough participants to establish a robust effect and this inflated the type I error. If this is the case, then the effect of context and skewness is weaker than we initially expected, and much of the observed variation comes down to individual participant pairs.

Individual variation suggests some participants are more likely to underspecify than others, and this is independent of our experimental manipulations. It is well documented that individuals vary on a whole host of measures, such as attention span and memory (Sauce & Matzel, 2013), and this might impact upon whether or not they learn and use a compositional language. Some participants in Shape-Different were already predicted to maintain a compositional language if they were exposed to a Shape-Skewed meaning space (conflicting biases). This leaves us with 8 participant pairs in which we would predict underspecified systems due to a reinforcing bias: shape is the only useful dimension for discrimination and there are fewer shapes than colours. However, if individual variation does play a big role in whether participants underspecify or not, then it may have been the case that the participants who were more likely to underspecify found themselves in the Both-Different and Colour-Different conditions in Experiment 3.

One way to resolve this issue is to collect larger samples. This allows us to detect whether there is an effect of our manipulations to context as well as to clarify the size of this effect. A drop in the effect size, as a result of a larger sample, suggests we are dealing with a small number bias (Tversky & Kahneman, 1971): the tendency for small datasets to show a strong effect that decreases with larger, more representative samples. One possibility, which we attempted

to investigate in this experiment, is that certain context-types increase the probability of participants underspecifying. If this is the case, and the effect size is small, then participants in Shape-Different and Colour-Different are predicted to underspecify more often than participants in Both-Different. Importantly, this small effect is mediated by individual variation: some participants will maintain and generalise a compositional system independent of which condition they find themselves in.

Another possibility is that we observe patterns of underspecification independent of condition. Such a result would bolster the case for individual variation being the main driver of the results in experiments 2 and 3. Individual variation relates to the prior probability of participants underspecifying or generalising the compositional system, with some participants having a higher probability of underspecifying. The prediction here is that underspecification should occur at similar levels in all three conditions. Under this account, context is still predicted to play a role, except now it is not a motivation for underspecification; instead, it is a communicative constraint on how participants underspecify, i.e., the dimension participants choose to encode and to not encode for conveying the intended meaning. So, when underspecification does occur, we predict that participants in Shape-Different only encode shape in their linguistic systems, participants in Colour-Different only encode colour in their linguistic system, and Both-Different will have some systems where only shape is encoded and some systems where only colour is encoded.

Individual variation is also potentially problematic with regards to the complexities of communication and interaction. For instance, we do not know how individual variation to learn a compositional system impacts upon strategy choice in dyadic interaction. It might be the case that, within a dyad, one participant (A) learns and generalises the compositional mapping and another participant (B) learns and generalises an underspecified mapping. This then introduces four possible outcomes in the experiment: participant A accommodates towards participant B (resulting in an underspecified system), participant B accommodates toward participant A (resulting in a compositional system), participants A and B accommodate towards one another (resulting in a noisy system with some compositional mappings and some underspecified mappings) or both participants are egocentric and maintain their own systems (resulting in two distinct idiolects, where participant A's productions are compositional and participant B's productions are underspecified). As such, we are now dealing with two levels of individual variation, with the first being variation in the capacity to learn a compositional mapping given exposure to a subset of possible form-meaning mappings, and the

second being variation in strategy choice during communication (e.g., whether participants are egocentric or accommodating).

3.4 Design Issues and Future Directions

There are a number of areas where the design of the experiment could be improved and extended. Three areas of immediate relevance are *the meaning space*, *the referential context*, and *the task*.

3.4.1 Changes to the meaning space

Perhaps the most egregious mistake was to change the colour, and by proxy the brightness, of the stimuli in the second experiment. Colour saliency can be manipulated based on brightness and this might partly explain why participants in the second experiment were more likely to generalise and encode colour. One obvious way to control for this confound is to systematically manipulate the appearance parameters (hue, lightness, brightness, chroma, colourfulness, and saturation) of the stimuli – and see whether certain parameters increase the probability of participants underspecifying. Another approach would be to manipulate the saliency of shape. For instance, the bias for shape can be weakened and strengthened with subtle manipulations to the stimuli, with one relevant manipulation being to remove the faces of the stimuli (as objects with eyes dampen the shape bias, see Jones, Smith & Landau, 1991).

A related problem is that the manipulations to skewness were not comprehensive, as we did not investigate a balanced meaning space (i.e., one where there was an equal number of shapes and colours). The rationale for manipulating skewness, rather than introducing a balanced meaning space, is that this is the simplest manipulation: we wanted to know whether having more colour features relative to shape features (colour-skewness) was creating a preference to encode shape. Introducing a counterbalance, whereby half of the participant pairs were exposed to a meaning space where there were more shape features relative to colour features (shape-skewness), allowed us to explicitly test this hypothesis. A balanced meaning space does not test this claim as it introduces an additional confound in that the size of the meaning space changes (it either shrinks to 3x3 or increases to 4x4). Future work needs to more carefully consider how manipulations to the size, complexity and skewness of the meaning space influence the loss of compositionality.

3.4.2 Changes to the context

When it comes to context, the first issue that needs to be addressed is methodological: Is the Both-Different condition a good baseline? Our initial assumption was that participants in Both-Different were more likely to maintain a compositional language: as neither shape nor colour were backgrounded it was hypothesised that this is easier for participants to learn the compositional mapping (when compared with Shape-Different and Colour-Different where one of the dimensions was backgrounded and not relevant for discrimination). However, this assumption ultimately rests on the dimensions being equal, without there being a preference to encode one over another. Both Experiment 2 and Experiment 3 had biased meaning spaces (as discussed above): to truly test these manipulations we would need a balanced meaning space.

The context used in this experiment was relatively simple: it consisted of one target and one distractor with manipulations being to whether or not a feature dimension was relevant to discrimination. Considerable work still needs to be done in establishing the effect size of manipulations to the contextual parameters (e.g., size, variability, complexity). Recent studies have shown that manipulations to visual-contextual factors, such as the number of dimensions on which objects in a scene differ (*scene variation*: Koolen *et al.*, 2013) and the colour variability of referents in a context (*polychrome* versus *monochrome*) influence the production of colour adjectives (Rubio-Fernández, 2016). Furthermore, there are conceptual questions about the validity of these manipulations, and whether or not they actually capture context in the sense used by linguists. For a start, the present experiment uses an external, referential context as a substitute for processes that normally take place in the discursive or syntactic contexts. Future experiments need to establish whether these differences are important when operationalising variables.

3.4.3 Changes to the task

It is still an open question as to the locus of innovation and change in the loss of compositionality. Our approach is consistent with claims that these changes initially arise from imperfect inference (e.g., Andersen, 1973; Lightfoot, 1979; Kuteva, 2001; Kiparsky, 2012). Under this account, the learner does not infer the underlying compositional rule, and instead opts for a non-compositional interpretation. The important point is that the learner simply fails to infer the original mapping. For our experiment, the imperfect inference takes place during the learning phase, with the change then becoming apparent in the communica-

tion task where participants need to generalise to an unseen set of meanings. If future work wants to focus on imperfect inference, then it might be better placed to simplify the task structure and have a discrimination game where participants learn and reproduce a language (without a communication component).

This account can be contrasted with the pragmatic reasoning perspective (e.g., Nerlich & Clarke, 1992; Sperber & Wilson, 1995; Traugott & Dasher, 2002; Smith & Hoefler, 2015): here, the locus of innovation and change is the speaker, who strategically uses a compositional form in a novel way that obscures the original mapping. If pragmatic reasoning is the main mechanism, one possible avenue of exploration is to manipulate how the communication task is framed: one framing could emphasise communicative success (as in the current experiment) whereas another framing might emphasise efficiency (e.g., be as quick as possible) or creativity. As it currently stands, our current set up is poorly equipped to distinguish between these two accounts: even though it is more consistent with imperfect inference, we simply do not know whether participants are failing to learn the compositional mapping or whether they are instead choosing to underspecify (as this is pragmatically economical).

3.5 Conclusion

This chapter began with the observation that the loss of compositionality appears to be motivated by context. To test this claim we set out to answer the following question: Do manipulations to context disrupt an initially compositional language in a discrimination game? More specifically, do contexts that background a particular feature increase the probability of participants underspecifying? The results of the second experiment suggest that manipulations to context interact with a bias to encode shape. This meant participants were more likely to underspecify when shape was the relevant feature (i.e., Shape-Different and Both-Different). When shape was backgrounded participants were more likely to maintain a compositional system (i.e., Colour-Different).

A possible explanation for the shape bias is an experimental confound in the number of feature dimensions, i.e., there were more colours than shapes. Experiment 3 followed up on this possibility by explicitly manipulating the skewness in the number of features to test whether or not it interacted with our manipulations to context. The findings somewhat support skewness as a key factor in whether or not a language loses its compositionality. Furthermore, skewness does appear to interact with context: when there are fewer colour features we

find that participants underspecify on this dimension in the Colour-Different and Both-Different conditions (in contrast to Experiment 2). However, the picture is still far from clear, and there are several puzzles that remain open for future work. First of all, the overall effect size in the experiment decreased, with participants underspecifying less. Second, the shape bias completely disappeared, even in cases where it was predicted (i.e., in Shape-Different where there were fewer shapes than colours). Lastly, for those cases where our predictions matched up with the results, we still do not have a clear explanation for why participants underspecified in these instances and not others.

Overall, this leaves us in a situation with many unanswered questions. One of which is the extent that context can independently disrupt an initially compositional system (i.e., without the presence of other factors such as skewness). Future work can remedy such issues by simplifying the experimental design and using this as a baseline for exploring further manipulations.

Chapter 4

Signal autonomy is shaped by contextual predictability

4.1 Introduction

In his book, *Arenas of Language Use*, Herb Clark (1992) argues that speakers adhere to a *Principle of Optimal Design*:

The speaker designs his utterance in such a way that he has good reason to believe that the addressees can readily and uniquely compute what he meant [...] The Principle of Optimal Design relies crucially on the notion of *common ground*, technically the mutual knowledge, beliefs, and assumptions shared by the speaker and addressees... In our proposal, the speaker intends each addressee to base his inferences not on just *any* knowledge or beliefs he may have, but only on their *mutual* knowledge or beliefs – their common ground. (Clark, 1992: 80-81; emphases in original).

Clark recognised that language is a useful tool in establishing common ground through creating conventional form-meaning mappings shared between speakers and hearers (also see: Grice, 1957; Lewis, 1969; Freyd, 1983; Parikh, 2001). For a language to arrive at these conventional form-meaning mappings a tradeoff needs to be reached between what a speaker needs to express and the amount of inferential effort required by the hearer. Chapter 4 investigates this tension between speakers and hearers by directly manipulating *contextual predictability*: the extent to which a speaker can estimate, and therefore exploit, the contextual information that a hearer is likely to use in interpreting an utterance. In particular, we argue that contextual predictability is causally related to *signal autonomy*, defined as the degree to which a signal can be interpreted without recourse to contextual

information. Decreasing contextual predictability is therefore predicted to result in increases in signal autonomy.

Experiment 4 manipulates two aspects of the referential context in an asymmetric communication game (where speakers and hearers are assigned fixed roles): (i) whether or not a speaker has access to the contextual information (*Access to Context*); and (ii) the consistency with which a particular dimension (e.g., colour) is relevant in discrimination across successive trials (*Context-Type*). The results demonstrate that contextual predictability does shape the degree of signal autonomy: when the context is highly predictable, languages are organised to be less autonomous (more context-dependent) through combining linguistic signals with context to reduce uncertainty about the intended meaning. When the context decreases in predictability, speakers favour strategies that promote autonomous signals, allowing linguistic systems to reduce their context dependency.

4.2 Author contributions

The following section contains a paper which was co-authored with my supervisors, Simon Kirby and Kenny Smith, and submitted to *Cognition*. The experiments were conceived during supervision meetings, with both co-authors contributing to the analysis and writing of the paper.

4.3 Winters, Kirby & Smith (submitted): Signal autonomy is shaped by contextual predictability

Contextual predictability shapes signal autonomy

James Winters^{1,*}, Simon Kirby, Kenny Smith

Abstract

At the heart of human communication is the goal of reducing uncertainty about the intended meaning. This requires solving a *recurrent coordination problem* where speakers and hearers need to align on a shared system of communication. Governing the tradeoff between what a speaker needs to express and the amount of inferential effort required by hearer is *contextual predictability*: to what extent a speaker can estimate and therefore exploit the contextual information that a hearer is likely to use in interpreting an utterance. This relationship between context and communicative pressures has important consequences for how languages are structured. In this paper, we test the claim that contextual predictability is causally related to *signal autonomy*: the degree to which a signal can be interpreted in isolation, without recourse to contextual information. Using an asymmetric communication game, where speakers and hearers are assigned fixed roles, we test for the effect of contextual predictability on signal autonomy by manipulating two aspects of the referential context: (i) whether or not a speaker has access to the contextual information used by the hearer in interpreting their utterance; and (ii) the extent to which successful communication requires the encoding of a consistent set of semantic dimensions. Our results demonstrate that contextual predictability shapes the degree of signal autonomy: when the context is highly predictable (i.e., the speaker has access to the context in which their utterances will be interpreted, and the dimension which discriminates between meanings in context is consistent across communicative episodes), languages develop which rely heavily on the context to reduce uncertainty about the intended meaning. When the context is less predictable, speakers favour systems composed of autonomous signals, where all potentially

*Corresponding author
Preprint submitted to *Cognition*
Email address: winters@shh.mpg.de (James Winters)

October 10, 2016

relevant semantic dimensions are explicitly encoded. Taken together, these results suggest that pragmatic factors play a central role in shaping the linguistic systems that emerge over repeated interactions between speakers and hearers.

Keywords: Language evolution, interaction, communication games, context, pragmatics

1. Introduction

Reducing uncertainty about the intended meaning is fundamental to any good communication system (Piantadosi, Tily & Gibson, 2012; Ramscar & Port, 2015). In achieving this aim, speakers and hearers need to coordinate with one another, relying not only on the creation of conventional forms, but also on the way these forms interact with the contextual information at hand (Lewis, 1969; Clark, 1996; Croft, 2000; Sperber & Wilson, 2005; Scott-Phillips, 2015). Without context, linguistic systems such as English would be woefully ambiguous, leaving the sentence *She passed the mole* uninterpretable as to whether the verb *passed* refers to *a form of motion* or *an act of giving* and whether the noun *mole* refers to *a small burrowing mammal*, *a person engaged in espionage*, *a brand of Mexican sauce* or *a type of causeway*. In short, when the context is known and informative, it helps in reducing uncertainty (Piantadosi et al., 2012).

Context in this sense is the mutual cognitive environment in which an utterance is situated (Sperber & Wilson, 2005) and acts as a *frame of interpretation* (Goffman, 1974; Minsky, 1975; Fauconnier, 1985; Fillmore, 1985): determining what is and is not informative for reducing uncertainty in interpretation. This consists of a figure (the target of interpretation), a ground (the immediate information brought to the act of interpretation), and a background (prior knowledge derived from previous frames) (Duranti & Goodwin, 1992; Terkourafi, 2009). And, as with any environment, the context will vary: some contexts are regular and predictable, whereas others fluctuate and are unpredictable. When viewed in this way, the context is a variable that determines the extent to which a speaker can estimate, and therefore exploit, information that the hearer can use

to reduce uncertainty about the intended meaning – its *contextual predictability*.

For instance, if a speaker is providing directions to the nearest grocery store, then the context includes information in the immediate environment, such as the general direction of the store relative to the present position of the interlocutors, as well as background knowledge about how a hearer is likely to interpret an utterance given the outcomes of previous interactions. Predictable contexts are therefore those where the speaker is able to use information provided by the context to reduce uncertainty about their intended meaning for the hearer: if the grocery store is near a park, and the speaker and hearer share knowledge about where this park is located, then saying “there’s a grocery store about five minutes away, next to the park where we play rugby” is sufficient for the hearer to find the grocery store. This is in contrast to a situation where the speaker and hearer are strangers and uncertainty exists as to the knowledge they both share with one another (e.g., the hearer is a tourist and does not know about the existence of a nearby park).

This relationship between context, meaning and uncertainty leads to an interesting trade-off in how linguistic systems are organised. Languages vary in their degree of *signal autonomy*: “the capacity for an utterance to be interpreted in isolation, without recourse to implicit linguistic, cultural, contextual or cotextual knowledge. Non-autonomous expression combines linguistic signals with context, pragmatics, paralinguistic signals and the like” (Wray & Grace, 2007: 556). One hypothesis is that autonomy is favoured in situations where speakers and hearers cannot rely on context for disambiguation (Kay, 1977): autonomous signals are advantageous insofar as they reduce reliance on shared social and physical context in favour of internal structure (Snow et al., 1991: 90-91; Hurford, 2011).

In this paper we present experimental evidence demonstrating that the degree of signal autonomy is causally related to contextual predictability: in an experiment where participants interact using an artificial language, highly predictable contexts favour systems composed of non-autonomous, context-dependent signals, whereas decreasing contextual predictability results in in-

creased autonomy (context-independence). Crucially, these systems arise from the competing demands of speaker effort and hearer inference, with the degree of contextual predictability shaping the tradeoff between these two constraints.

1.1. Signal Autonomy and Contextual Predictability

No natural language has completely autonomous signals in the sense of unambiguous clarity; context is always involved in reducing uncertainty about the intended meaning. But it is relatively uncontroversial to say there are degrees of autonomy. Consider the possible use of referring expressions in Figure 1. Describing the object on the left in contexts A and B could be achieved with the referential expression *the metal cup* – this expression is capable of discriminating between referents in both contexts. Yet, based on a long history of psycholinguistic studies, it is only in context A where the expression *the metal cup* is used, with *the cup* being preferred when the adjective is not needed for discrimination (Olson, 1970; Pechmann, 1989; Sedivy, 2005; for review, see: Konopka & Brown-Schmidt, 2014)¹. Similar work in *audience design* (Clark, 1996) shows that speakers produce longer, more elaborate expressions when the hearer is perceived to be less knowledgeable about a topic (Brown & Dell, 1987; Isaacs & Clark, 1989; Horton & Gerrig, 2005; Heller et al., 2009) and the use of scalar-modified expressions is partially contingent on whether or not a speaker and a hearer share the same referential context (Keysar et al., 2000; Nadig & Sedivy, 2002).

¹The use of colour adjectives provides an interesting counterexample to this general picture. Unlike material and scalar adjectives, which tend to be dependent on context, colour adjectives are often used even when they are uninformative for discrimination (e.g., Sedivy, 2005; Arts et al., 2011). A growing body of work into these *Redundant Colour Adjectives* (RCA) provides two explanations (Rubio-Fernandez, 2016). First, the use of RCAs tends to be contingent on the semantic category, as evident in their presence for atypical objects (e.g., *the brown banana*) and clothes (e.g., collocations such as *black tie*) and their absence in typical (e.g., *the banana*) and geometrical figures (Dale & Reiter, 1995; Sedivy, 2003; Grodner & Sedivy, 2011). Second, speakers tend to produce RCAs when colour helps facilitate object recognition (e.g., polychrome versus monochrome displays), as well as when the language uses pre-nominal (e.g., English) as opposed to post-nominal adjectives (e.g., Spanish) (Rubio-Fernandez, 2016).



Figure 1: In context A, an English speaker can discriminate between both objects by using *the cup* or *the bowl*, whereas in context B they must use more elaborate expressions: *the metal cup* and *the wooden cup* (assuming the speaker obeys the rules of English for adjective use).

In fact, our everyday language use is littered with options for more or less autonomous expressions. Contrast the use of indexical (context-dependent) and non-indexical (autonomous) forms of language: when referring to the day after today, English users will tend to say *tomorrow*, rather than the more autonomous counterpart of a specific date (e.g., *July 5th 2016*) (Hurford, 2011). Both are perfectly valid forms of expressing the relevant meaning, yet indexical forms are preferred in the presence of shared-knowledge and predictable contexts (e.g., James lives on *this* street), whereas non-indexical forms are useful in providing specific information in the absence of such contexts (e.g., James lives on Milton Street). The key point is that all these examples vary in their contextual predictability: the extent to which a speaker exploits contextual information to reduce uncertainty about the intended meaning for the hearer.

Contextual predictability, then, can be seen as an organising principle for how we use language: when the context is highly predictable, use less au-

onomous forms, and in less predictable contexts use more autonomous forms. However, differences in autonomy are not just found in how we use language in context, but also in the way language is structured. For instance, contrast nouns with concrete, physical senses (e.g., *dog* and *computer*) and nouns that are considered maximally vague (e.g., *stuff* and *thing*): whereas *computer* is relatively autonomous, in that we can get some sense of the intended meaning out of context, *thing* needs contextual enrichment to even get an approximation as to what a speaker might be referring². This difference in autonomy holds across parts of speech categories as well: the most frequent nouns show a tendency to be more autonomous than the most common verbs and adjectives (Engelkamp, Zimmer & Mohr, 1990: 190). And, at even higher level of organisation, grammatical morphemes are less autonomous than lexical morphemes, with the former being more bound to sentence context (Mihatsch, 2009).

Explanations for language change also bind together these two concepts of context and autonomy. Work in *grammaticalization* provides a useful illustration of the interaction between contextual predictability and autonomy across historical timescales (Hopper & Traugott, 2003; Traugott & Trousdale, 2012). A classic example of grammaticalization is the development of *gonna* from *be going to*. In the time of Shakespeare’s English, *be going to* had its literal meaning of a subject travelling to a location in order to do something, yet over time this construction extended its original meaning to also include instances where the motion verb (*go*) and the purpose clause (*to* + infinitive) came to express intentionality and future possibility (Hopper & Traugott, 1993; Croft, 2000). By expanding its range of uses, *be going to* underwent a decrease in autonomy, and became increasingly reliant on contextual information for disambiguation. For example, saying *the leaves are going to fall off the tree* is unambiguously referring to a near future event, with contextual information guiding the reader

² *Thing* is an example of vagueness in language, defined as the modification of a linguistic item, phrase or utterance to make its meaning less precise (Channell, 1994: 20; Cutting 2007). In this sense, an individual may not intend to refer to a specific referent, with vagueness being “stretched and negotiated to suit the moment-to-moment communicative needs” (Zhang, 2011: 573).

toward the intended meaning. But there are situations where *be going to* is ambiguous as to whether it refers to a form of motion or future intent (e.g., *I am going to take a nap*) – and this opens up the possibility of low contextual predictability triggering uncertainty in interpretation. One solution to the problem of ambiguity is to enrich the interpretation with additional linguistic information, as evident in *Maria is going to go to London*, where both future intent and motion are separately expressed (making the construction more autonomous). Another way is to create a more autonomous form, as is the case with *gonna*, which unambiguously refers to near future events (e.g., *The leaves are gonna fall off the tree* is grammatical whereas this is not the case for **Maria is gonna London*).

Languages also appear to vary in their degree of autonomy (Kay, 1977; Wray & Grace, 2007; Hurford, 2011). An extreme example of the cross-linguistic variation in autonomy is found in Riau Indonesian – a colloquial variety of Malay/Indonesian with minimal syntactic structure and highly context-dependent expressions (for review see Gil, 2005). For instance, consider the possible interpretations found in combining the forms *ayam* (“chicken”) and *makan* (“eat”): *ayam makan* or *makan ayam* can refer to anything from *the chicken is eating* to *the chickens are eating* or *someone is eating the chicken* or even *someone is eating with the chicken* (Gil, 2005; Hurford, 2011). In short, the phrase *ayam makan* or *makan ayam* involves anything to do with chicken and eating; contextual information and hearer inference do the rest of the work in sifting through possible interpretations.

Riau Indonesian fits nicely into the picture presented by Wray & Grace (2007) that situations characterised by high degrees of shared knowledge result in less autonomous signals. The reason for this is that shared knowledge increases contextual predictability: if the speaker and hearer have common knowledge about a particular topic, and this knowledge is adequate for distinguishing between possible interpretations, then there is no need to explicitly express such information in the linguistic system.

There is therefore converging evidence from usage, change and typology

suggesting that differences in contextual predictability lead to differences in autonomy: high contextual predictability biases signals toward becoming less autonomous, whereas low contextual predictability is associated with an increase in signal autonomy. However, this still leaves us with the question: What are the underlying mechanisms linking contextual predictability with signal autonomy? Wray & Grace (2007: 556) offer one answer:

Individual utterances in any language will, of course, score differently on the autonomy scale, since it is part of the speaker's job to judge how much knowledge the hearer shares, and thus what it is appropriate not to mention in the interests of relevance and brevity.

As the above quote suggests, differences in autonomy appear to be related to competing motivations, which, in the next section, we argue consist of a pressure to reduce speaker effort and a pressure to minimise hearer uncertainty.

1.2. Competing Demands: Minimising Effort and Reducing Uncertainty

Connecting the relationship between autonomy and contextual predictability to underlying mechanisms requires we take seriously the *problem of linkage* (Kirby, 1999; 2012): How do the behaviours of individuals give rise to the particular structural properties of language? One solution to the problem of linkage is to consider how short-term strategies used in solving immediate communicative needs can give rise to language systems through long-term patterns of learning and use (Evans & Green, 2006; Steels, 2012; Beuls & Steels, 2013; Winters, Kirby & Smith, 2015; Pleyer & Winters, 2015).

Shaping these short-term strategies are the competing motivations of speakers and hearers (Zipf, 1949; Horn, 1993; Nettle, 1999; Frank & Goodman, 2012; Piantadosi et al., 2012). Here, a balance needs to be reached between the demands of the speaker, where the goal is to reduce effort in production, and the demands of the hearer, where the goal is to reduce effort in comprehension. Resolving these two forces require speakers and hearers to align on a system: one that reaches a tradeoff between minimising the energetic expense of the

speaker whilst also reducing uncertainty about the intended meaning for the hearer (Piantadosi et al., 2012). The core principle being that a speaker aims to be efficient and informative given the hearer’s common knowledge, context, and the task at hand (Frank & Goodman, 2012).

The pressures imposed during communication provide a clear prediction linking together signal autonomy and context. Autonomy is expected when contextual predictability is low. This is because speakers cannot rely on contextual information when conveying their intended meaning to the hearer, and must instead explicitly encode more information in the signal (increasing the probability of the hearer arriving at the correct interpretation). By contrast, when contextual predictability is high, speakers are expected to use less autonomous signals: a speaker can rely on context to get some meaning across, allowing them to reduce their overall effort and encode less information explicitly (without negatively impinging upon the hearer arriving at the correct interpretation).

An optimal system therefore minimises both speaker effort and hearer inference, resulting in a low cost and functionally adequate mode of communication. A simple example of this is found in research relating word length, ambiguity and predictability given the context (Cohen Priva, 2008; Tily & Piantadosi, 2009; Piantadosi et al., 2011; Rohde et al., 2012; Mahowald et al., 2013; Seyfarth, 2014). For instance, Tily & Piantadosi (2009) show that pronouns are used for referents which are more predictable in context, whereas proper names and full NPs are reserved for less predictable contexts. Even in pairs of near synonymous forms, such as *exam* and *examination*, the short forms tend to have a lower information content than the long forms when controlling for meaning (Mahowald et al., 2012). All of this points to the general conclusion that “shorter and less informative expressions are favoured when less information is sufficient to carry the message” (Tily & Piantadosi, 2009: 1).

Such results lend weight to the idea that contextual predictability governs the extent to which a speaker can reduce their effort without significantly taxing hearer inference. But there are important theoretical differences in how this tradeoff is reached. Levinson (2000: 29), for instance, argues that an optimal

system should be skewed in favour of reducing speaker effort at the expense of minimising hearer uncertainty, his rationale being that “inference is cheap, articulation expensive, and thus the design requirements are for a system that maximizes inference”. Meanwhile, audience design (e.g., Bell, 1984; Clark & Murphy, 1982) and accommodation theory (Giles, Coupland & Coupland, 1991) paint a slightly different picture: here, perspective taking mechanisms are emphasised, with knowledge about hearers being used to optimise messages (Clark & Carlson, 1982; Levelt, 1989; Blokpoel et al., 2012).

A middle ground between these two perspectives is found in *probabilistic pragmatics* (Franke & Degen, 2015; Franke & Jäger, 2016; Pfeifer, 2016) which argues that population-level phenomena are shaped by heterogeneous pragmatic reasoners. That is, individuals do not necessarily enter into conversations with the same default assumptions, and are most likely updating their beliefs about the pragmatic behaviour of others as the conversation progresses. Solving this *recurrent coordination problem* requires the use of communicative strategies to build common ground – the central task being to negotiate a shared system that allows for successful communication between speakers and hearers (Lewis, 1969; Freyd, 1983; Sperber & Wilson, 1995; Clark, 1996; Croft, 2000; Gärdenfors, 2000; Parikh, 2001).

Language games (Wittgenstein, 1953; Steels, 1999: 197-8) are useful for viewing these pragmatic strategies in action. Figure 2 shows a simple language game where the speaker has to convey the same intended meaning in two different contexts (A and B). Context A has referents that share the same colour but differ in shape, and context B has referents that share the same shape but differ in colour. In the shared context, the speaker and hearer both have access to the same contextual information, whereas in the unshared context the speaker only sees the referent they need to convey. Assuming speakers have knowledge about a set of signals, and are free to combine them, then there are three possibilities (for the target referent): *the blue one*, *the square*, *the blue square*. The optimal solution in the shared context is to minimise effort and only convey what is necessary: use *the square* in context A and *the blue one* in context B.

This is because the context is to some extent predictable – the speaker and the hearer share knowledge about the relevant distinctions – and speakers need not expend effort by using a more autonomous form to achieve communicative success. However, in the unshared context, speakers only have access to the target they need to convey, and are therefore unable to condition their signals on contextual information. To ensure communicative success, speakers need to expend effort and specify both dimensions in a single signal. By using the more autonomous form, *the blue square*, speakers can be sure to convey their intended meaning across both contexts.

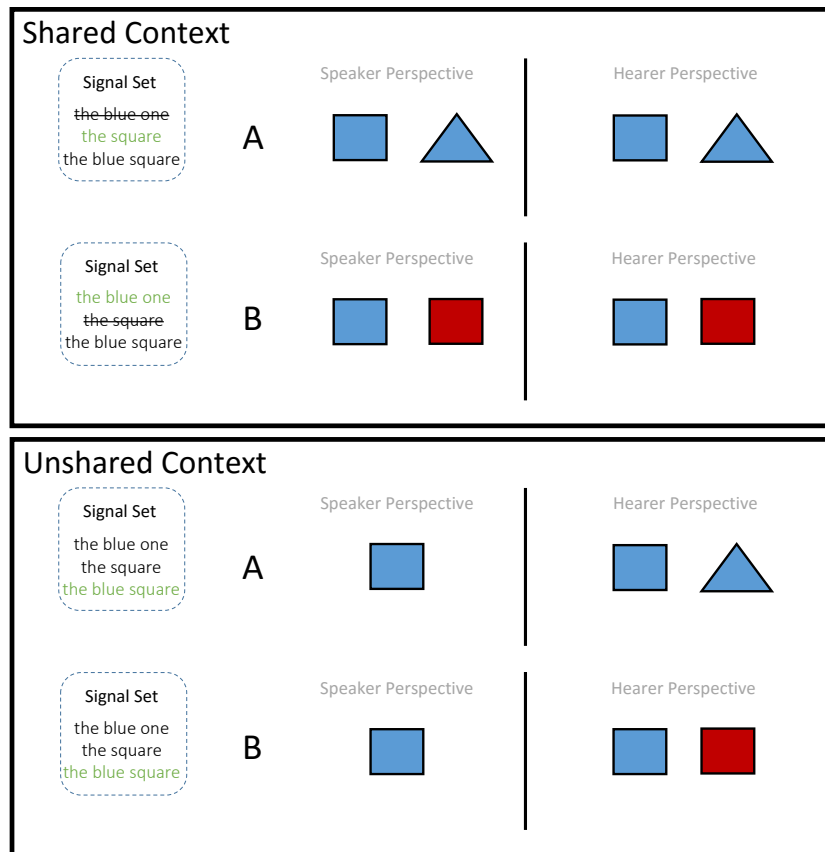


Figure 2: An example of a language game where the speaker has a set of signals for conveying a particular meaning (in this case, a blue square). There are two contexts: A and B. In context A, the target image shares the same colour as its distractor (but differs in shape), whereas in context B the target image shares the same shape as its distractor (but differs in shape). For the Shared Context (top box), the speaker has access to the same contextual information as the hearer. By contrast, in the Unshared Context (bottom box), the speaker does not have access to the same contextual information as the hearer. The green coloured signals are the preferred signals for communicating an intended meaning in a particular scenario. The signals with a strikethrough are those which will not convey the intended meaning in a particular scenario.

The imagined English-speaking interlocutors in Figure 2 already have a fully functioning communication system in place for expressing the features of a tar-

get referent. In order to study how such communication systems develop, and how the form they take is modulated by contextual predictability, the present experiment uses an artificial language paradigm (e.g., Saffran, Aslin & Newport, 1996; Kirby, Cornish & Smith, 2008). In particular, we explore how short-term strategies of achieving communicative success in a referential context influence the emergence of language systems – and therefore solve the problem of linkage for signal autonomy.

1.3. Access to Context and Context-Type

For our present purposes, we focus on the emergence of two types of language system: *context-dependent* and *autonomous*. Context-dependent systems are expected when signals can be conditioned on contextual information in order to reduce effort. Out of context, signals in a context-dependent system are less capable of distinguishing between referents, and we therefore expect that these systems will only emerge when the context is to some extent predictable. Autonomous systems are expected to develop when the context is less predictable and more information needs to be explicitly encoded in the linguistic signal.

Communication games are a useful tool for exploring the dynamics underpinning communicative behaviour in a laboratory setting (for review, see: Galantucci, Garrod & Roberts, 2012; Tamariz & Kirby, 2016). In our experiment, participants are first trained on an initially ambiguous artificial language, and then placed in an *asymmetric communication game* (Moreno & Baggio, 2014; Nowak & Baggio, 2016) where they are assigned fixed roles as either speaker or hearer. Speaker and hearer play a series of guessing games (Steels, 2003; Silvey, Kirby & Smith, 2015; Winters, Kirby & Smith, 2015): the task is for the hearer to discriminate between a target object and a set of distractor objects using a signal provided by the speaker. Possible referents were drawn from a set of images which vary in shape and colour.

Our rationale for using asymmetric participant roles is to explicitly investigate how contextual predictability influences a speaker’s behaviour. If participants alternate between the role of speaker and hearer, then a speaker at time t

will have gained some relevant knowledge about what contextual information a hearer has at $t+1$, which will diminish the effect of our manipulations (see below) — fixing the participant roles for the duration of the experiment removes this possibility. We also train our participants pre-interaction on an ambiguous language which underspecifies whether labels encode shape, colour or both (see below for details of how this is achieved). This creates a *generalisation pressure*, allowing us to explore how speakers convey novel meanings, and how much information they choose to encode explicitly in the linguistic signal.

To test for the effect of contextual predictability on autonomy, we made two manipulations: (i) Access to context (Shared Context/Unshared Context), and (ii) Context-type (Shape-Different/Mixed). In the *Shared Context* conditions, speakers have access to the context against which their utterance will be interpreted (i.e., the array of target and distractors that the hearer is confronted with), whereas in the *Unshared Context* conditions speakers only see the target in isolation (although the speaker’s task remains the same: to produce a signal which allows the hearer to distinguish a target from a set of distractors). Speakers in Shared Context conditions therefore have knowledge about what distinctions they need to make on a trial-by-trial basis, whereas speakers in Unshared Context conditions only know what target they need to convey (without any contextual information about the context against which their utterance will be interpreted, and therefore what the relevant distinctions are for the hearer in a particular trial).

Our second manipulation of contextual predictability involves context type: to what extent is a particular dimension (e.g., shape) relevant for discrimination across successive trials? For the *Shape-Different* conditions, the context-type remains consistent across trials, with targets and distractors always differing in shape (but sharing the same colour). *Mixed* conditions vary their context-type across trials: half of the trials consist of contexts in which the target and distractors differ in shape (but share the same colour) and half in which they differ in colour (but share the same shape). In Shape-Different conditions, encoding shape is therefore always sufficient to allow the hearer to retrieve the intended

meanings (although whether the speaker can see this directly or must infer it from the pattern of communicative successes and failures will depend on the Shared vs Unshared manipulation); in Mixed conditions, some trials will require the encoding of shape, some will require colour to be encoded (and whether or not the speaker knows which dimension is relevant for a given trial will again depend on the Shared/Unshared manipulation). This gives us four conditions: *Shape-Different + Shared Context*, *Shape-Different + Unshared Context*, *Mixed + Shared Context*, *Mixed + Unshared Context*.

In terms of contextual predictability, the Shape-Different + Shared Context condition is the most predictable both within and across trials: the context-type is consistent, in that Shape is always the relevant feature for discrimination, and the speaker has access to the same contextual information as the hearer. The optimal solution here is for speakers to use the contextual information to generalise and only encode shape in their signals, resulting in a system with low signal autonomy (out of context a signal has a decreased capacity to discriminate between referents). This is an example of underspecification: speakers are able to reduce effort because they can abstract across referents, using a single signal to refer to different meanings based on a shared feature (Silvey, Kirby & Smith, 2015; Winters, Kirby & Smith, 2015). Underspecified systems are communicatively functional if contextual information aids in discriminating between meanings: information about the context, when combined with the interpretive clues provided by the signal, allows a hearer to correctly infer the intended meaning.

On the opposite end of the scale of contextual predictability is the Mixed + Unshared Context condition: context-type varies between trials where objects in the context differ in shape (but share the same colour) and trials where objects in the context differ in colour (but share the same shape), with access to this contextual information being unavailable for the speaker (they only ever see the target that needs to be conveyed). This low contextual predictability means that underspecified systems will be ineffective — in order to be sure of conveying their intended meaning, speakers must instead employ strategies that increase

signal autonomy, e.g., by encoding both shape and colour on every trial.

For the Shape-Different + Unshared Context and Mixed + Shared Context conditions there is one manipulation which decreases contextual predictability and another which increases contextual predictability. In the Shape-Different + Unshared Context condition, the fact that the speaker lacks access to the context favours strategies that increase signal autonomy, as the speaker has no contextual information regarding what distinctions they need to convey. However, the across-trial predictability potentially allows speakers to reduce their signal autonomy, as encoding shape in the linguistic is always sufficient for conveying the intended meaning. Whether or not a speaker opts for strategies that increase or decrease signal autonomy is somewhat contingent on the initial assumptions a speaker brings to the task as well as the feedback they receive from hearers. The first option is for the speaker to prioritise reducing their effort and only convey shape, but this strategy runs the risk of failing to successfully communicate with the hearer (remember: the speaker does not know which distinctions are relevant for conveying the intended meaning). The second option is to increase the signal autonomy: this is advantageous in that autonomous signals are less reliant on the external context for disambiguation. However, there is a cost associated with this strategy, as speakers now need to encode more specific information into the signal.

A similar story holds for the Mixed + Shared Context condition. This time the variability across trials decreases contextual predictability, as the context-type varies between trials where objects in the context differ in colour and trials where objects in the context differ in shape; however, the fact that the speaker has access to the same contextual information as the hearer should increase contextual predictability. Again, there are two viable strategies in this condition. The first is a context-dependent strategy: speakers underspecify and encode shape and colour in distinct signals, with the interpretation being conditioned on the context-type. As an example, imagine conveying a blue rectangle in a shape-different and a colour-different context: for the shape-different context, where the set of distractors consist of a blue blob, a blue oval, and a blue star,

the speaker uses a signal that conveys rectangle, whereas the speaker uses a signal to convey blue in the colour-different context, where the set of distractors consist of a grey rectangle, a red rectangle, and a yellow rectangle. This strategy is useful inasmuch as the speaker is able to reduce their effort – a set of eight signals is sufficient to convey all 16 meanings. Crucially, the burden is mostly shifted onto the hearer, who must figure out how the signal-meaning mappings vary according to context-type. The second strategy is for speakers to provide more autonomous signals. But, as mentioned above, autonomous signals come with additional costs of encoding more specific information into a given signal.

2. Method

2.1. Participants

120 undergraduate and graduate students at the University of Edinburgh (79 female, 41 male, median age 20) were recruited via the Student And Graduate Employment database and randomly assigned to one of the four possible conditions (see § 2.3.3). Each condition consisted of a pair of participants who learned an artificial language (see § 2.2) and then used this language in a communication game (see § 2.3.2). Participants were paid £5 for their participation.

2.2. Stimuli: Images and Target Language

Participants were asked to learn and then use an ‘alien language’, consisting of lower-case labels paired with images. There were 16 images that varied along three features: shape, colour and a unique identifier (see figure 3 for examples). Four of these 16 images were randomly selected for training, such that each colour and shape was represented exactly once and each of the four images therefore differed from all the others in both colour and shape. Each image was then assigned a label as follows: From a set of vowels (a,e,i,o,u) and consonants (g,h,k,l,m,n,p,w) we randomly generated nine CV syllables which were then used to randomly generate a set of four 2-3 syllable words. Since the four images used during training differed in both shape and colour, the training labels in

this language were therefore ambiguous with respect to whether they referred to colour, to shape, or to both colour and shape (or, equivalently, the unique identifier).

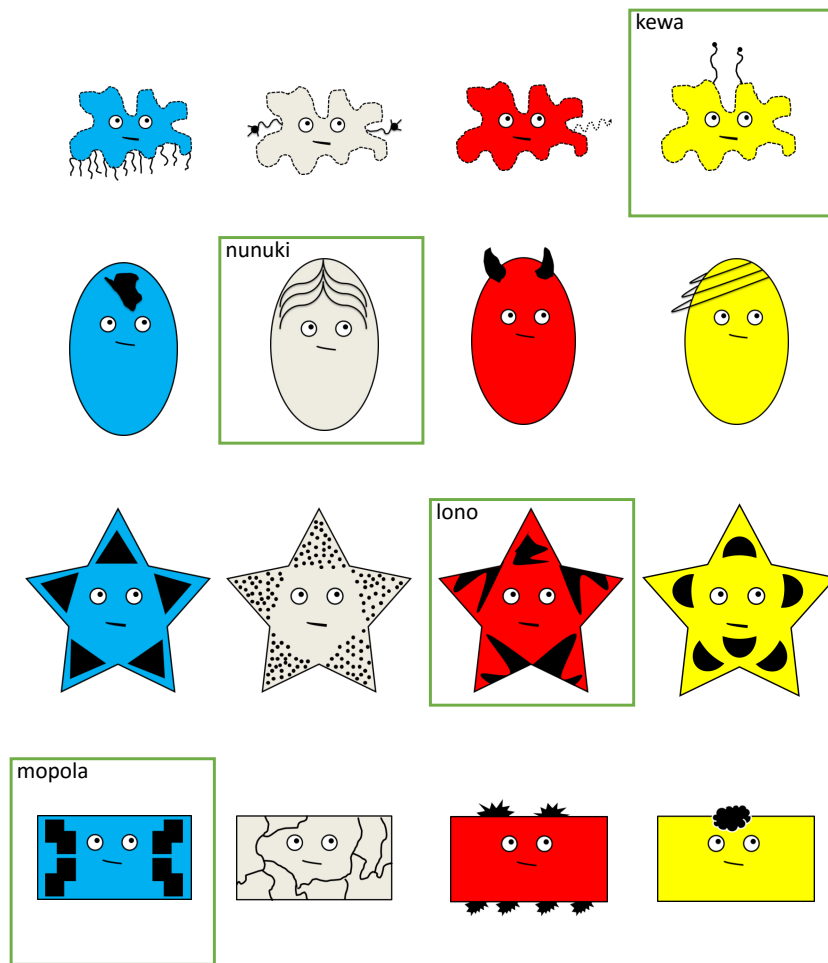


Figure 3: The meaning space used in our experiment. Referents vary along three dimensions: a shape, a colour, and a unique identifier. To create an initially ambiguous language we first randomly selected four meanings that differed from one another on both shape and colour. We then paired these four meanings with randomly generated signals (see signals and meanings inside green boxes).

2.3. Procedure

At the start of the experiment, participants were told they would first have to learn and then communicate using an alien language. Participants completed the experiment in separate booths on networked computers. The experiment consisted of two main phases: a *learning phase* and a *communication phase*. Before each phase began, participants were given detailed information on what that phase would involve and were explicitly told not to use English or any other language they knew during the experiment³. For the learning phase, participants were trained separately, and it was only during the communication phase that they interacted (remotely, over the computer network).

2.3.1. Learning Phase: Training and Testing

The learning phase was broken up into two components: a set of *training blocks* and a set of *testing blocks*. In each training trial, the participant was presented with an image and a label. After two seconds, the label would disappear, and the participant was then prompted to retype the label before proceeding to the next trial. Each training block consisted of twelve trials (each of the four training images was seen three times, with the order of trials randomised within the block). For the testing trials, the participant was presented with an image and prompted to type the label that corresponded to the image. Once they typed the label, the participant was given feedback as to whether or not they were correct – if incorrect, they were shown the correct label before moving onto the next trial. A testing block consisted of sixteen trials (meaning that each of the four training images was seen four times, with the order of trials randomised within the block). The training and testing blocks were interleaved as follows: 2 training blocks, 1 testing block, 2 training blocks, 1 testing block.

³Compliance with the instructions was excellent – there were no cases in which participants used English or any other natural language in the current experiment

2.3.2. *Communication Phase*

During the communication phase of the experiment, participants within a pair were randomly assigned fixed roles of either *speaker* or *hearer*:

Speaker On each communication trial, the speaker was shown a target image that was highlighted with a green border. Whether or not a speaker could view the distractors depended on the experimental condition (see § 2.3.3). The speaker was then prompted to type a description that would best communicate the target to the matcher. Speakers were free to type any description they wished (subject to the requirement to avoid English or any other natural language). This description was then sent to the hearer’s computer.

Hearer Hearers were presented with an array of four images; the description provided by the director appearing underneath. Of these four images, one was the target image and the other three images were distractors. Distractors were randomly generated within the constraints imposed by the experimental conditions (see § 2.3.3). The hearer’s goal was to click on the image they thought corresponded to the description provided.

Participants were tested on all 16 images during the communication phase, requiring the speaker to generalise from the signals provided for the four images in the training set, and the hearer to interpret these generalisations. Following each trial, both speaker and hearer were given feedback as to whether or not the hearer had correctly identified the target image described by the speaker: both participants were simply informed whether the hearer was correct or incorrect. The communication phase was comprised of three blocks, with each block consisting of 32 trials (trial order was randomised and each of the 16 images appeared as the target image twice within a block).

2.3.3. *Manipulating Access to Context and Context-Type*

During communication, we manipulated two variables associated with the referential context: (i) access to the referential context; (ii) the referential con-

text type.

The first manipulation consisted of whether or not a director had access to the referential context that the matcher saw. In the *Shared Context* conditions, directors were exposed to the same referential context as the matcher: that is, they had access to an array consisting of the target and its distractors. Conversely, for the *Unshared Context* conditions, directors only had access to the target image.

A second manipulation was made to the possible combinations of target and distractor images within a single trial. For the *Mixed* conditions, half of all trials consist of referential contexts in which the target and its distractors have different shapes (but share the same colour) and half in which they have different colours (but share the same shape). For the *Shape-Different* conditions, the referential context-type remains consistent across trials, with the target and distractors having different shapes (but sharing the same colour). This gives us four conditions: (i) *Shape-Different + Shared Context*; (ii) *Shape-Different + Unshared Context*; (iii) *Mixed + Shared Context*; (iv) *Mixed + Unshared Context* (see Figure 2).

2.4. Dependent Variables and Hypotheses

2.4.1. Communicative success

To measure communicative success we recorded the number of successful trials between the speaker and hearer, i.e., when the hearer clicked on the target image. The maximum success score was 96 points for three blocks of 32 trials. The purpose of this measure is to see whether the communication systems which develop during interaction are communicatively functional.

2.4.2. Total number of signals

One way to distinguish between autonomous and context-dependent systems is count the total number of unique signals produced. To convey all 16 meanings an autonomous system requires more unique signals (16) than a context-dependent system (where the same signal can be reused to express different

meanings).

2.4.3. Measuring uncertainty: conditional entropy

To quantify the types of mappings between signals and meanings we measure the *conditional entropy* (Shannon, 1948) of meanings given signals for the speaker’s productions during interaction (Winters, Kirby & Smith, 2014). This gives us a measure of predictability that can be applied to meaning uncertainty. $H(M|S)$ is the expected entropy (i.e., uncertainty) over meanings given a signal, and therefore captures meaning uncertainty,

$$H(M|S) = - \sum_{s \in S} P(s) \sum_{m \in M} P(m|s) \log P(m|s) \quad (1)$$

where the rightmost sum is the entropy over meanings given a particular signal $s \in S$. $P(m|s)$ is the probability that meaning m is the intended meaning given that signal s has been produced, and $P(s)$ is the probability that signal s will be produced (for any meaning). A high $H(M|S)$ corresponds to low signal autonomy, i.e., out of context a signal is highly uninformative about the intended meaning, with a speaker reusing that same signal to convey several meanings. By contrast, an autonomous signal should have zero $H(M|S)$, as each signal a speaker uses only conveys one meaning.

While this measure captures the extent to which signals are autonomous, it does not distinguish between context-dependent and counter-functional ambiguity. For context-dependent ambiguity, contextual information contributes to reducing uncertainty about the intended meaning, whereas with counter-functional ambiguity this is not the case. To differentiate these two possibilities we also include a measure of meaning uncertainty in context, $H(M|S, C)$:

$$H(M|S, C) = - \sum_{s, c \in S, C} P(s, c) \sum_{m \in M} P(m|s, c) \log P(m|s, c) \quad (2)$$

The rightmost sum now takes into account the entropy over meanings given a particular signal in context $s, c \in S, C$. A context $c \in C$ is an array of four meanings taken from set M and is constructed so that each meaning shares one

feature in common and differs on the other feature, e.g., *shape-different blue* = {*blue blob, blue oval, blue square, blue star*} and *colour-different star* = {*blue star, grey star, red star, yellow star*}.

Our general prediction is that, even though systems will vary in $H(M|S)$ due to the effects of contextual predictability on autonomy, all systems will gradually decrease their $H(M|S, C)$ over time. This is expected when languages are adapting to optimise their communicative success. As such, a context-dependent system should therefore have high $H(M|S)$ but low $H(M|S, C)$: this difference between $H(M|S)$ and $H(M|S, C)$ indicates that signals are hard to interpret in isolation but contextual information helps in identifying the intended meaning, i.e., the communication system is functionally adequate in context. Conversely, for an autonomous system, we expect both the $H(M|S)$ and the $H(M|S, C)$ to be low, i.e., signals are informative even out of context.

2.4.4. Hypotheses

Here we provide a set of hypotheses and predictions related to our specific measurements:

Hypothesis One: Communication systems will be functionally adequate for identifying the intended meaning in context. As such, we predict all conditions will reach a communicative success score higher than chance (>25%).

Hypothesis Two: Speakers will consistently use contextual information to reduce their effort and only specify shape in the Shape-Different + Shared Context condition (low autonomy). This will result in underspecified systems with a low number of signals, a high $H(M|S)$ and a low $H(M|S, C)$.

Hypothesis Three: Strategy choices in the Shape-Different + Unshared Context condition and the Mixed + Shared Context condition will depend on the relative weighting participants give to minimising speaker effort, hearer effort and accuracy of communication. If a speaker prioritises reducing their own effort, then the resulting systems will underspecify and

only encode shape (low autonomy); reflected in a low number of signals, a high $H(M|S)$ and a low $H(M|S, C)$. However, if a speaker prioritises reducing hearer uncertainty, then the resulting systems will be more autonomous and have a greater number of signals, a low $H(M|S)$ and a low $H(M|S, C)$.

Hypothesis Four: Speakers will consistently converge on strategies that promote autonomous signals in the Mixed + Unshared Context condition. This will result in systems with a greater number of signals, a low $H(M|S)$ and a low $H(M|S, C)$.

3. Results

Our analyses involved four separate mixed effect models (*lme4*: Bates, Machler, Bolker & Walker, 2014) based on the dependent variables of (a) communicative success, (b) number of unique signals, (c) $H(M|S)$, and (d) $H(M|S, C)$. For communicative success we used a logistic mixed effect model and for number of unique signals, $H(M|S)$, and $H(M|S, C)$ we used linear mixed effect models. Context-Type (Shape-Different or Mixed), Access to Context (Shared Context or Unshared Context)⁴ and Block (1, 2 and 3 — Block was coded such that model intercepts give performance at block 1) were entered as fixed effects with interactions. We included random intercepts for Participant and initial training language, and random slopes for all fixed effects and associated interactions (following the *keep it maximal* approach: Barr et al., 2013). P-values for the fixed effects in the linear mixed effect model were obtained using the *lmerTest* package (Kuznetsova, Brockhoff & Christensen, 2016).

3.1. Communicative success

All conditions show levels of communicative accuracy substantially higher than chance ($> 25\%$) in communicating with one another. This is confirmed by a

⁴Context-Type and Access to Context were centered fixed effects, coded such that positive values of β indicate higher communicative success/number of signals/entropy in Shape-Different or Shared conditions

logistic mixed effect model⁵, which has a significant intercept ($\beta = 3.440$, $SE = 0.230$, $p < .001$) indicating performance above chance. This model also indicates significant effects of Block ($\beta = 0.325$, $SE = 0.074$, $p < .001$): as can be seen in Figure 4, participants in all four conditions are increasing their success rate over time. Access to Context ($\beta = 1.301$, $SE = 0.401$, $p < .002$) and Context-Type ($\beta = 2.480$, $SE = 0.405$, $p < .001$) are also significant predictors of communicative success: conditions where the speaker has access to the hearer's context (Shared Context) and where the context-type remains stable across trials (Shape-Different) leads to higher communicative success (as highlighted by the positive coefficients for Access to Context and Context-Type). Finally, the Shape-Different + Shared Context condition (with the highest contextual predictability) is clearly something of an outlier, and the model indicates a significant Context-Type x Access to Context interaction ($\beta = 1.874$, $SE = 0.801$, $p = .019$) which shows that communicative success in this condition is higher than we would expect given the independent contributions of shared and stable context. There were no other significant main effects or interactions ($p > .103$).

⁵The model was adjusted to control for the chance level of communicative success, which in our case is 25% as there is 1 target and 3 distractors within a single trial.

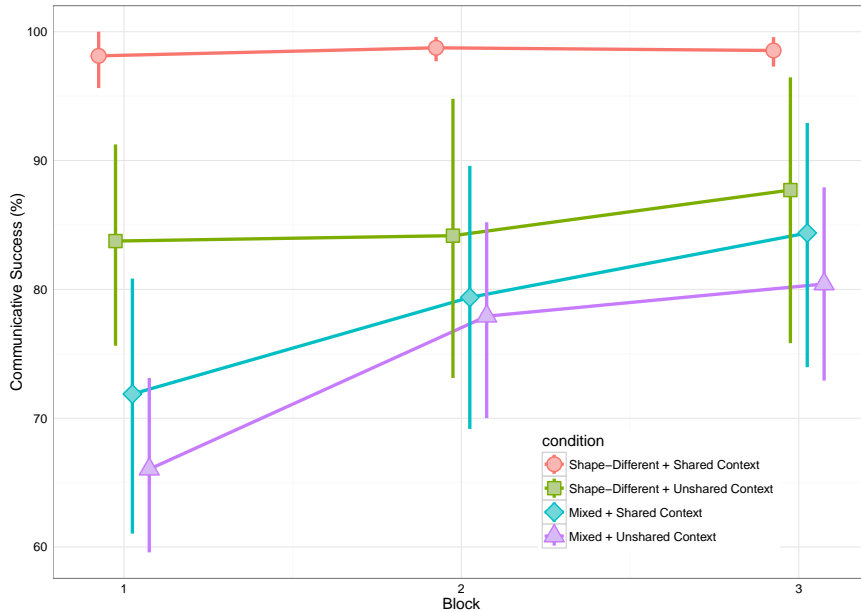


Figure 4: Mean communicative success score by condition over blocks 1-3. Error bars represent bootstrapped 95% CIs where the original data was resampled 10000 times using the *Hmisc package* (Harrell Jr et al., 2014). All of the conditions show a steady increase in communicative success (although performance in the Shape-Different + Shared Context condition is near ceiling from block 1), and all participants scored higher than chance ($> 25\%$) in selecting the target image, indicating that the systems are communicatively functional for identifying the intended meaning.

3.2. Number of Unique Signals

Figure 5 shows the number of unique signals across condition. As was the case for communicative success, there are significant effects of Context-Type ($\beta = -7.361$, $SE = 1.414$, $p < .001$), Access to Context ($\beta = -5.006$, $SE = 1.414$, $p < .001$), and a significant interaction for these two predictors (Context-Type x Access to Context: $\beta = -7.033$, $SE = 2.828$, $p = .015$): having a shared context and context-types which remain stable over time (i.e., Shape-Different conditions) are associated with smaller signal inventories, suggesting a lesser degree of signal autonomy, and the combination of these manipulations results in very small lexicons in the Shape-Different + Shared Context condition.

The marginally significant two-way interaction for Context-Type x Block ($\beta = -0.983$, $SE = 0.520$, $p = .061$) tells us that the average number of signals decreases over time in Shape-Different conditions; however, the significant three-way interaction between Access to Context, Context Type and Block ($\beta = 2.567$, $SE = 1.039$, $p = .015$) suggest that the number of signals in Shape-Different + Shared Context condition is trending upwards for some participant pairs (counteracting the overall negative effect of Block: $\beta = -0.358$, $SE = 0.230$, $p = .171$). Overall, these series of results suggest that conditions where the number of unique signals is low from the offset (i.e., Shape-Different + Shared Context at Block 1) show an increase in the number of signals, but this increase is relatively small when compared with the total number of unique signals in the other three conditions. All other predictors and associated interactions were non-significant ($p > .171$).

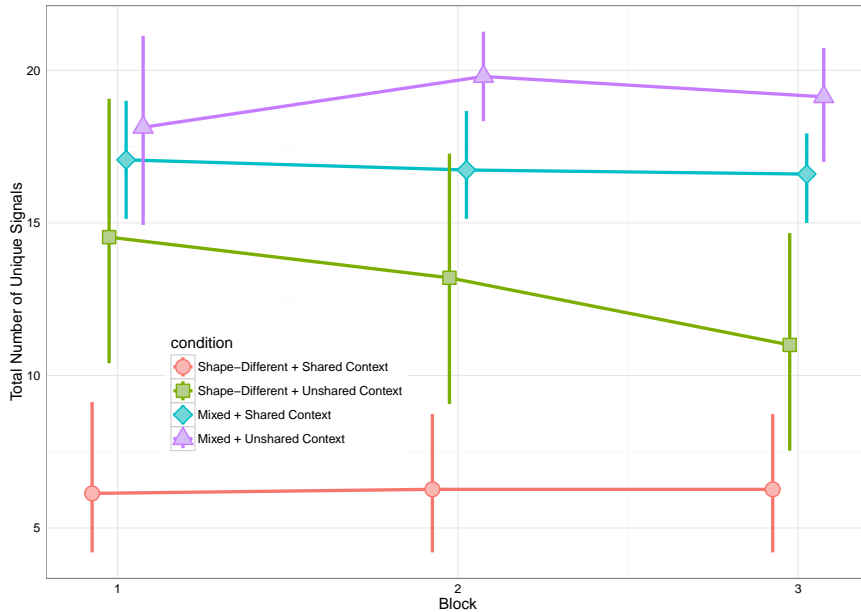


Figure 5: Mean number of unique signals by condition over blocks. Participants in Shape-Different + Shared Context condition use fewer unique signals than participants in the other conditions, and in the Shape-Different + Unshared Context condition the average number of unique signals tends to decrease over time, which suggests the systems are moving away from being autonomous by increasing their context-dependency. Participants in Shape-Different + Shared Context are near to optimal in using the minimal number of signals (4) required for successful communication in context, whereas participants in Mixed + Shared Context and Mixed + Unshared Context use roughly the number of signals (16) required in an autonomous signalling system. Signalling systems in the Shape-Different + Unshared Context condition appear to be non-autonomous, in that on average participants tend to produce fewer than 16 signals — the number of unique signals changes over the course of interaction. Error bars represent bootstrapped 95% confidence intervals.

3.3. Conditional Entropy: Meaning Uncertainty

Figure 6 plots the conditional entropy of meanings given signals, $H(M|S)$, against condition. As a visual inspection of the plot suggests, both Context-Type ($\beta = 0.783$, $SE = 0.154$, $p < .001$) and Access to Context ($\beta = 0.406$, $SE = 0.154$, $p = .011$) are significant predictors of $H(M|S)$: when contextual predictability increases so too does the out-of-context ambiguity as measured

by $H(M|S)$, indicating that higher contextual predictability leads to lower signal autonomy. Again, the significant interaction between these two predictors (Context-Type x Access to Context: $\beta = 0.756$, $SE = 0.296$, $p = .013$) indicates that the combination of a shared and stable context in the Shape-Different + Shared condition produces systems of even lower autonomy than we would expect through the independent contributions of either factor alone. Finally, there is a significant interaction between Context-Type and Block ($\beta = 0.119$, $SE = 0.025$, $p < .001$) which shows that Context-Type influences the evolution of the signalling systems over successive blocks: Shape-Different conditions tend to produce higher $H(M|S)$, with participants in Shape-Different + Unshared increasing their average $H(M|S)$, whereas participants in Mixed conditions show a steady *decrease*. No other interactions were significant ($p > .117$).

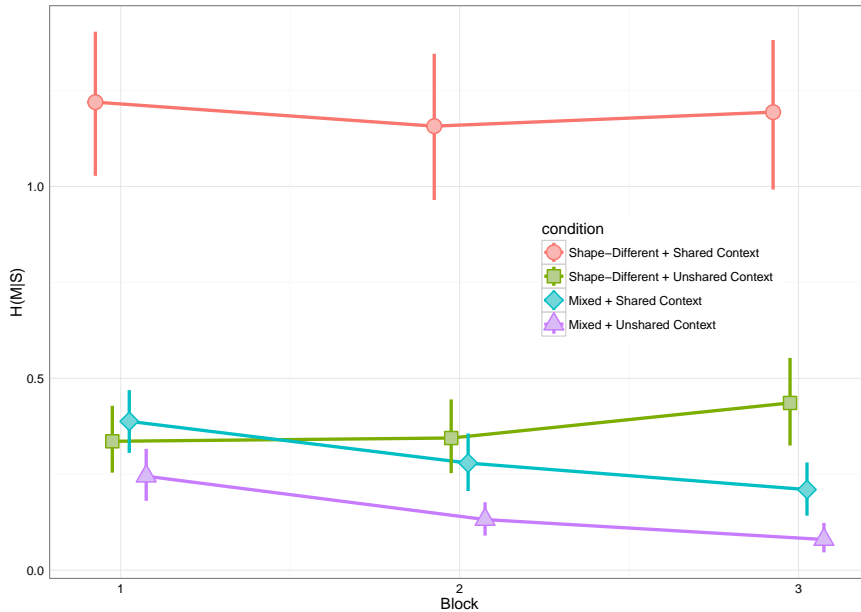


Figure 6: Mean meaning uncertainty, measured as $H(M|S)$, by condition over blocks 1-3. Higher entropy values indicate greater out-of-context ambiguity, i.e., lower signal autonomy. Whereas participants in Mixed conditions tend to decrease their $H(M|S)$ over time, those in Shape-Different conditions tend to either have a high $H(M|S)$ throughout (as in Shape-Different + Shared Context) or gradually increase their $H(M|S)$ (as in Shape-Different + Unshared Context). The higher $H(M|S)$ in Shape-Different + Shared Context suggests participants in this condition are reusing the same signals to express multiple meanings. The low $H(M|S)$ for Mixed + Unshared Context indicates that participants in this condition use each signal to express fewer meanings, i.e., are producing signals which are unambiguous even out of context.

3.4. Conditional Entropy: Meaning Uncertainty in Context

For the conditional entropy of meanings given signals in context, $H(M|S, C)$, the statistical analysis reveals unexpected differences between conditions: that is, conditions where the speaker does have access to the hearer’s context (Shared) and the context-type consistently discriminates on the basis of shape (Shape-Different) are, on average, more likely to produce languages with lower levels of uncertainty about the intended meaning in context (Access to Context: $\beta = -0.051$, $SE = 0.019$, $p = .011$; Context-Type: $\beta = -0.058$, $SE = 0.019$,

$p = .003$). When contextual predictability is high, this allows systems to get close to an optimal configuration for communication, with many of systems in Shape-Different + Shared Context reaching zero entropy (i.e., no uncertainty about the intended meaning in context). For the other three conditions, where contextual predictability is lower, systems tend to be suboptimal (see figure 7). But it is important to note that the low entropy values in all four conditions tells us, even at block 1, all of the communication systems are relatively good at identifying the intended meaning in context.

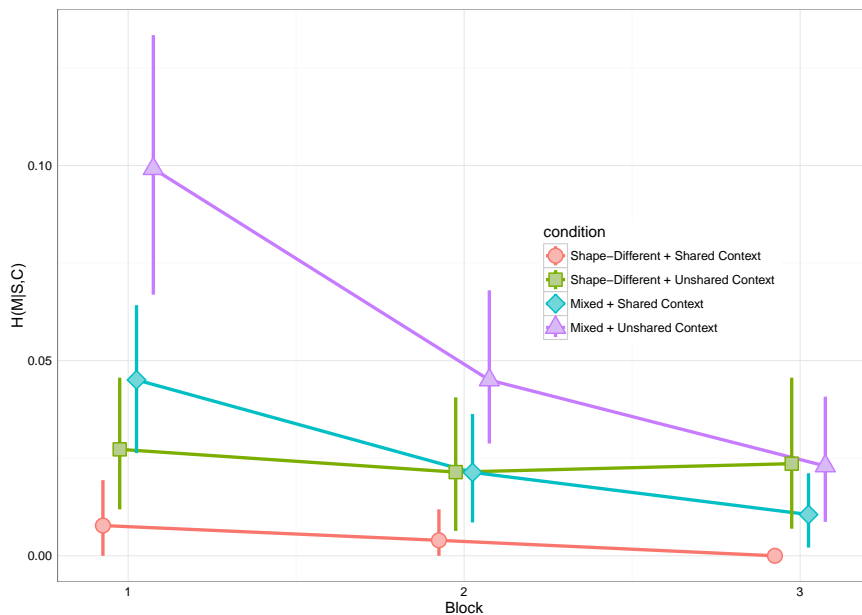


Figure 7: Degree of meaning uncertainty, measured as $H(M|S,C)$, against Condition over blocks. Higher entropy values indicate a higher degree of ambiguity in context. The error bars represent bootstrapped 95% CIs. All conditions tend to decrease their $H(M|S,C)$ over successive blocks. This suggests the systems used by participants are moving toward an optimal configuration for identifying the intended meaning in context.

There was also a significant effect of Block ($\beta = -0.021$, $SE = 0.003$, $p < .001$), indicating that average entropy across all 4 conditions decreased over repeated communication; however, this was moderated by significant two-

and three-way interactions between Block, Context-Type and Access to Context. The significant interaction between Block and Context-Type ($\beta = 0.027$, $SE = 0.007$, $p < .001$), of the same magnitude as the simple effect of Block, suggests that the Shape-Different conditions (which start with low ambiguity) show relatively little decrease in ambiguity over blocks - the marginal three-way interaction between Context-Type, Access to Context and Block ($\beta = -0.026$, $SE = 0.014$, $p = .059$) suggests that ambiguity in the Shape-Different + Shared Context condition might be trending downwards rather than being flat, but not convincingly so. In contrast to the Shape-Different conditions, the Mixed conditions show a substantial decrease in ambiguity over blocks; however, the significant interaction between Block and Access to Context ($\beta = 0.017$, $SE = 0.007$, $p = .012$) indicates that effect is less pronounced in the Shared Context condition. Overall, this pattern of interactions suggests that those conditions in which ambiguity is low right from block 1 (most obviously the Shape-Different conditions) show relatively little subsequent decrease in ambiguity (since the systems are unambiguous from very early on); the conditions where ambiguity is higher at block 1 (both Mixed conditions) show larger decreases, most obviously in the Mixed + Unshared Context condition where ambiguity is highest at block 1 but drops very rapidly over the course of interaction to reach levels similar to those seen in the other conditions by block 3. The remaining interaction for Access to Context x Context-Type ($\beta = 0.041$, $SE = 0.038$, $p = .283$) is non-significant.

4. Discussion

We put forward the general hypothesis that contextual predictability shapes the degree of signal autonomy. To test this claim we manipulated both the speaker's ability to access the context in which their utterances were interpreted and the variability of context-types across trials. When the context is predictable, speakers organise languages to be less autonomous (more context-dependent), exploiting contextual information to reduce effort in production

while at the same time keeping uncertainty in comprehension low. In conditions with lower contextual predictability, speakers use more autonomous signals, and rely less on contextual information to discriminate between possible meanings. In line with previous work, these results demonstrate that languages adapt to their contextual niche (Piantadosi, Tily & Gibson, 2012; Silvey, Kirby & Smith, 2015; Winters, Kirby & Smith, 2015).

The key finding is that number of unique signals used and $H(M|S)$ are predicted by both Context-Type and Access to Context. Furthermore, even though $H(M|S)$ varies substantially between conditions, all of the communication systems which develop during interaction are communicatively functional (i.e., capable of discriminating between meanings in context), as indexed by our measures of communicative accuracy and $H(M|S, C)$. If the Context-Type is stable for discrimination across trials (Shape-Different), and speakers have access to this contextual information (Shared Context), then participants will produce higher $H(M|S)$ than participants in conditions where one or both of these variables is less predictable. Having highly predictable contextual information allows speakers to produce signals which map onto multiple meanings: both the signal and the context relay to the hearer what is and is not informative for discriminating between meanings. As the contextual predictability decreases, the speaker is unable to estimate, and therefore exploit, this contextual information, resulting in an increased pressure to create autonomous signals (i.e., signals which are identifiable out of context).

Feedback also plays an increasingly important role as contextual predictability decreases. For the Shape-Different + Shared Context condition, which was maximally predictable in terms of context-type and access to context, communicative success remained constant and high across blocks. This is because the majority of systems were context-dependent, allowing participants to use a small number of signals to quickly achieve a high success score: speakers used their knowledge of the context to leave out the colour-dimension, as this was irrelevant to communicative success, and only conveyed the shape-dimension in the linguistic system (see Figure 8 for example language). By contrast, in the Mixed

+ Unshared Context condition, which was maximally unpredictable in terms of context-type and access to context, the communicative success score gradually increased across blocks: here, autonomous systems emerged, with speakers specifying both colour and shape within the linguistic system (see Figure 9).

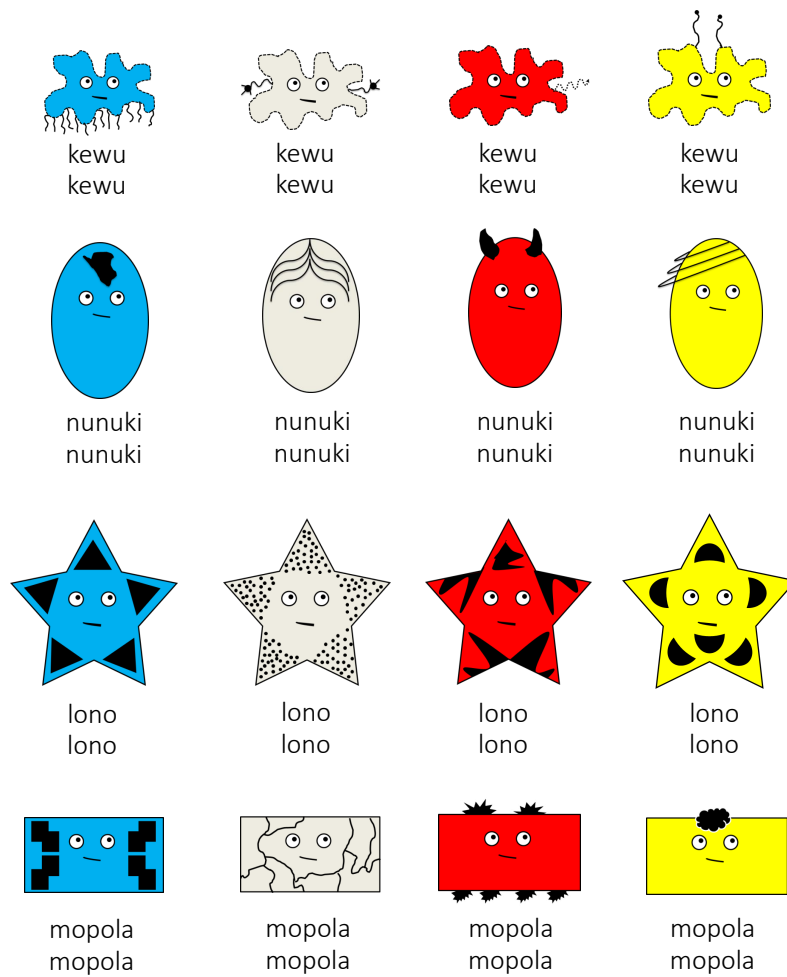


Figure 8: An example language from the final block of the Shape Different + Shared Context condition. The speaker labelled each meaning twice during interaction, both labels are shown here. In this case, the participants maintained the original four labels they were trained on, generalising them to only encode information about the shape dimension.

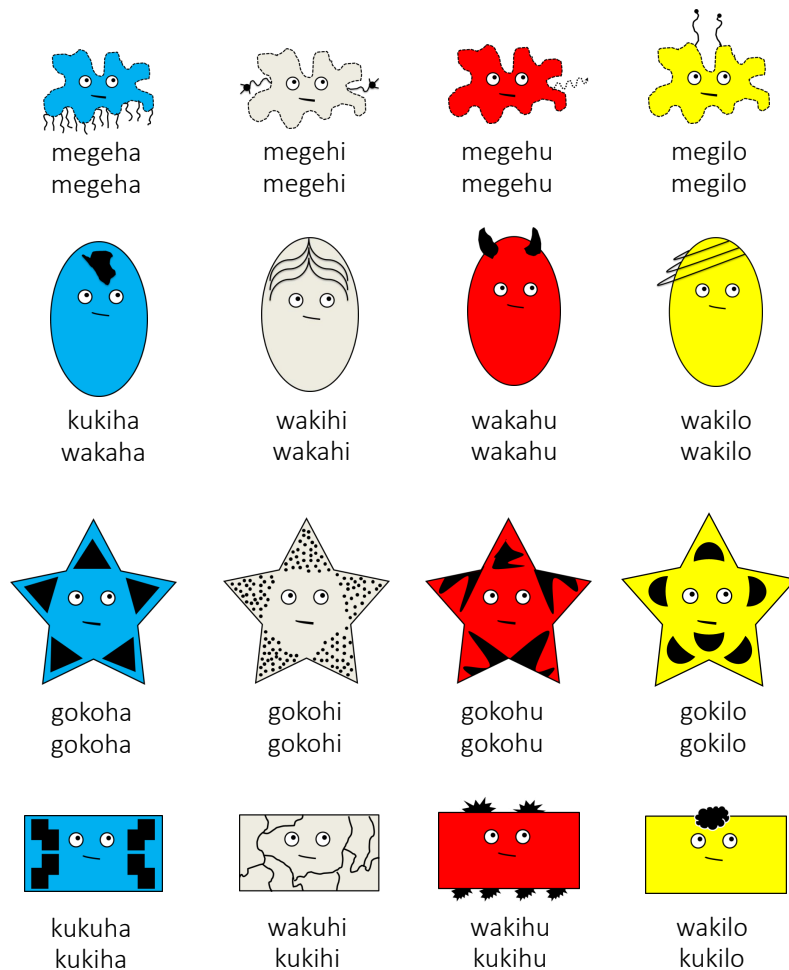


Figure 9: An example language from the final block of the Mixed + Unshared Context condition. Under each meaning are two signals arranged according to trial order (i.e., the first signal was produced prior to the second signal for conveying a particular meaning). Although there is some variability, the system can be described as compositional, where the meaning of an expression is a function of the meaning of its parts and the way in which these parts are combined (Partee, 1984: 281), with a single signal containing two component parts: the initial component identifies shape and the final component identifies colour. For example, if the initial component is *mege*, then it refers to a blob, and if the final syllable is *ha* it refers to blue. The combination of these component parts results in the signal *megeha*, meaning blue blob.

For participants in Shape-Different + Unshared Context, a gradual increase in communicative success is associated with a drop in the total number of unique signals, and an increase in $H(M|S)$: on average, speakers end up using fewer signals to convey more meanings, resulting in increasingly context-dependent systems (this is illustrated in Figure 10 where part of a signalling system becomes more context-dependent over successive blocks). The opposite is true for participants in Mixed + Shared Context: the average $H(M|S)$ decreases as the communicative success increases, with the set of signal-meaning mappings transitioning from a one-to-many to a one-to-one mapping (see Figure 11 for an example where part of a system becomes less context-dependent over successive blocks). The divergence between these two conditions suggests context-type exerts a stronger effect than access to context on the types of systems which emerge. Communicatively successful systems in Shape-Different + Unshared Context evolve to become increasingly context-dependent (and move closer to the systems found in Shape-Different + Shared Context) whereas the systems in Mixed + Shared Context become more autonomous (and move closer to the systems found in Mixed + Unshared Context).


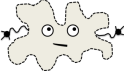

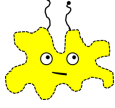
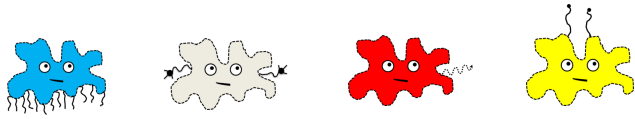
				
block 1	kim kim	lunimugi kimunus	kimunus kinumu	hamoho kimugi
block 2	kim kim	kinumus kinumus	kinumu kinumus	kinumus kimunugi
block 3	kim kim	kinumus kinumus	kinumus kinumus	kinumus kinumus

Figure 10: A subset of a gradually emerging context-dependent system in the Shape-Different + Unshared Context condition. At the first block, the system is relatively autonomous, in that each blob has a unique signal that identifies whether it is a blue, grey, red, or yellow one. However, by block 3 the signal *kinumus* refers to the grey, red, and yellow blobs (the exception being the blue blob which retains a unique identifying signal, *kim*). For the full language see supplementary materials.



block 1	memuno memunopewa	memuno memunonoka	memuno memunomemuno	memuno memunokihimo
block 2	memunopewa memunopewa	memunonoka memunonoka	memuno memunomemuno	memunokihimo memunokihimo
block 3	memunopewa memunopewa	memunonoka memunonoka	memuno memunomemuno	memunokihimo memunokihimo

Figure 11: A subset of a language in the Mixed + Shared Context condition showing the emergence of an autonomous system. The first row in each block contains signals used in shape-different contexts and the second row contains signals used in colour-different contexts. At block 1, the system is context-dependent, with an underspecified signal, *memuno*, being used in shape-different contexts to refer to blob, and compositional signals, *memunopewa*, *memunonoka*, *memunomemuno*, and *memunokihimo*, being employed for colour-different contexts. By block 3 the underspecified mappings are no longer used and the system is no longer context-dependent (compositional mappings are used in both shape-different and colour-different contexts). The only exception is for the red blob, which uses *memuno* in the shape-different context and a reduplicated form, *memunomemuno*, in the colour-different context. For the full language see supplementary materials.

The final systems (those at the last block of communication) in Shape-Different + Unshared Context and Mixed + Shared Context conditions tended to be more heterogeneous than those in conditions at the extremes of contextual predictability (Shape-Different + Shared Context, Mixed + Unshared Context). This suggests that differences in strategy choice at the individual level can result in markedly different systems of communication. For instance, if speakers prioritised reducing their effort, then the resulting systems have lower levels of autonomy, with the set of signals being conditioned on context-type. Conversely, if speakers aimed to reduce inference for the hearer, then systems at the final block were more autonomous. This variation in individuals lends weight to the idea of populations being composed of heterogeneous pragmatic reasoners;

speakers do not necessarily start out with the same initial assumptions, even if they eventually converge on the same set of behaviours (Franke & Degen, 2015). Our approach shows how to unmask these differences by manipulating whether contextual predictability is *reinforcing* (i.e., both manipulations either decrease or increase predictability) or *conflicting* (i.e., when one manipulation increases predictability and the other decreases predictability). When contextual predictability is reinforcing, as is the case in Shape-Different + Shared Context and Mixed + Unshared Context, participants are more likely to converge on similar systems. In cases where the contextual predictability is conflicting, as in Shape-Different + Unshared Context and Mixed + Shared Context, participants with differing initial assumptions can produce radically different systems of communication.

This variation in the emergence of context-dependent and autonomous systems is nicely demonstrated in the Mixed + Shared Context condition. Here, the systems could broadly be described as compositional, where the meaning of an expression is a function of the meaning of its parts and the way in which these parts are combined (Partee, 1984: 281). However, there are clear differences in autonomy for these systems: if you placed compositional systems on a cline of more to less autonomous, those found in the Mixed + Shared Context can generally be described as less autonomous than those found in the Mixed + Unshared Context Condition.

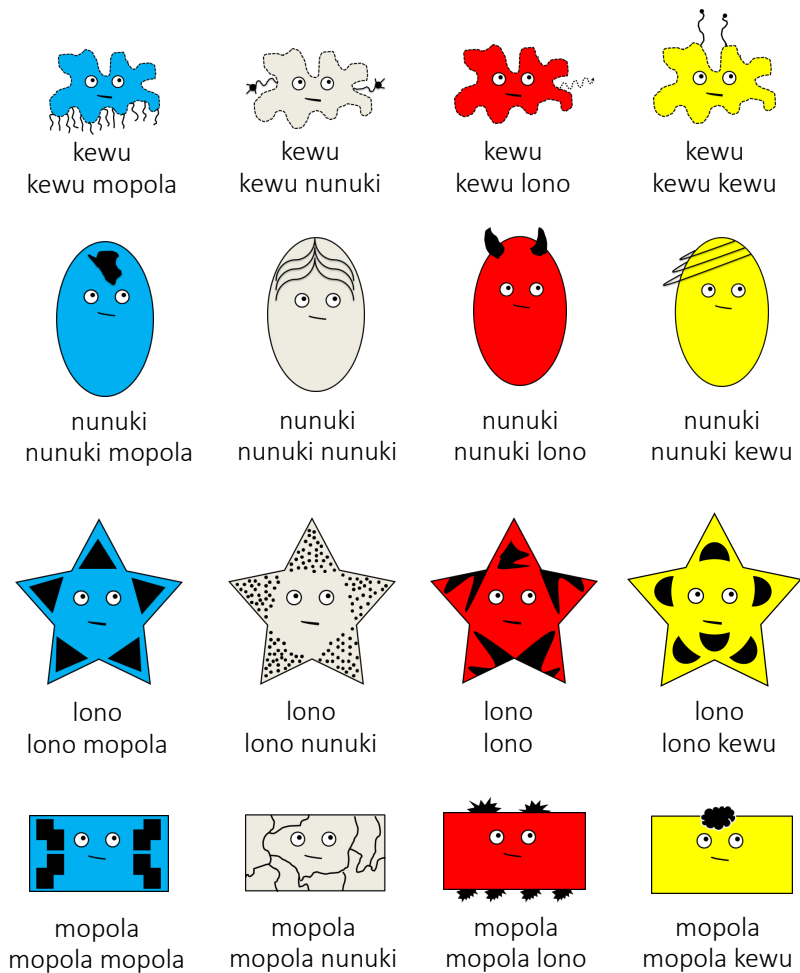


Figure 12: An example language from the final block of the Mixed Shared condition. Under each meaning are two signals: the first corresponds to the signal used in Shape-Different context-types and the second is the signal used in Colour-Different context-types. Notice that in the Shape-Different context-types an underspecified signal is used, whereas in Colour-Different context-types a compositional signal is used (where information about both colour and shape dimensions is encoded through the use of word order). For example, using the underspecified signal, *mopola*, in shape-different contexts refers to rectangle, whereas in colour-different contexts each meaning has a unique compositional signal (*mopola mopola*, *mopola nunuki*, *mopola lono*, *mopola kewu*). Interestingly, the compositional system is order-dependent, with a word initial *mopola* referring to rectangle, whereas a word final *mopola* refers to the colour blue.

Some systems in the Mixed + Shared Context condition are ones which produce marked and unmarked forms depending on the context (see Figure 12). For instance, an unmarked signal, *kewu*, is used to convey shape information (blob) in shape-different contexts, whereas in colour-different contexts marked forms are used (e.g., *kewu mopola* for the referent blob blue). This use of marked and unmarked forms speaks to Horn's (1993: 40-41) *Pragmatic Division of Labour*:

[...] given two co-extensive expressions, the more specialized form – briefer and/or more lexicalized – will tend to become R-associated [reduce speaker effort] with a particular unmarked, stereotypical meaning, use, or situation, while the use of the periphrastic or less lexicalized expression, typically (but not always) linguistically more complex or prolix, will tend to be Q-restricted [minimise hearer uncertainty] to those situations outside the stereotype, for which unmarked expression could not have been used appropriately.

Why would unmarked forms be associated with Shape-Different contexts and marked forms Colour-Different contexts? One possible explanation is a shape-bias (Diesendruck & Bloom, 2003): speakers reason that hearers are going to more easily guess the intended meaning in Shape-Different contexts than Colour-Different contexts. This could also be a low-level bias, in that speakers automatically associate their training forms with shape, and ignore colour. However, given the high level of lexical ambiguity in the use of these forms (e.g., *kewu* refers to the shape blob and the colour yellow depending on the context), a low-level bias cannot be the complete story.

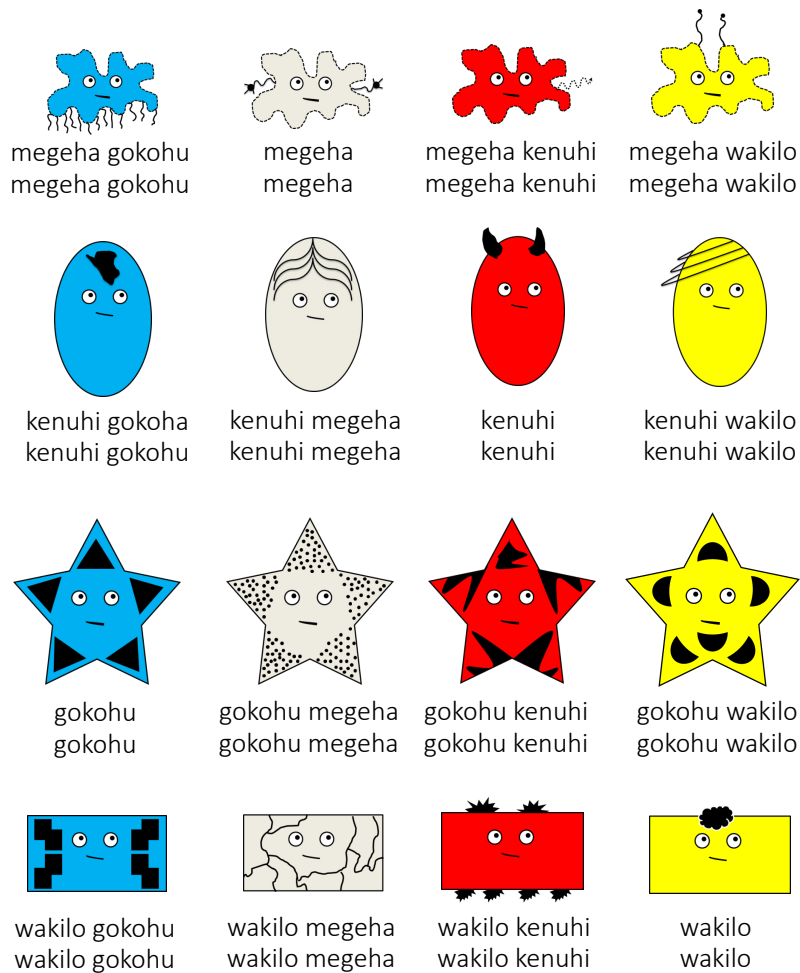


Figure 13: An example language from the final block of the Mixed Shared Context condition. Under each meaning are two signals: the first corresponds to the signal used in Shape-Different context-types and the second is the signal used in Colour-Different context-types. A single signal is composed of two parts: the initial signal refers to shape (e.g., *wakilo* refers to rectangle) and the second signal refers to colour (e.g., *wakilo* refers to yellow), regardless of context type.

In Mixed + Shared Context we also saw compositional systems where signals were not context-dependent (see Figure 13). While this system exhibits some ambiguity at the lexical level (*gokohu* can be used to convey shape or

colour), complete utterances are interpretable out of context and therefore autonomous, as each compositionally-formed signal unambiguously conveys the intended meaning. This process of shifting autonomy to different levels of organisation relates to grammaticalization, with lexical forms becoming increasingly dependent on the constraints of the linguistic system (Haspelmath, 2004).

Context-dependent compositionality also explains why participants in Mixed + Shared Context sometimes produce more signals than necessary for conveying all 16 referents: there are a set of autonomous signals which convey the intended meaning in colour-different contexts and a set of underspecified signals which convey the intended meaning in shape-different contexts. However, this fails to explain the even higher levels of variability in Mixed Unshared Context, where it is impossible for speakers to condition their signals on contextual information. A contributing factor to this unexpected signal variation is that speakers adjust their signals based on negative feedback (i.e., when the hearer clicks on the incorrect meaning). That is, if a speaker receives negative feedback for a particular signal-meaning mapping, then they are more likely to modify this signal for future communicative interactions (see Figure 14). This suggests (some) speakers are principally focused on the immediate communicative requirements; using feedback to fine-tune their signals as the interaction progresses. An additional advantage of this explanation is that it accounts for high signal variability at later stages of the experiment (where one might have expected more regular one-to-one mappings between signals and meanings): speakers and hearers get trapped in a coordination problem, where a speaker changes a signal (for a particular meaning) based on an incorrect guess by a hearer, and the hearer guesses incorrectly when presented with a different a signal (as they make the (approximate) inference that a modified signal must refer to a different meaning).


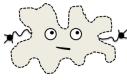










Speaker's Signal	Speaker's Intended Meaning	Hearer's Interpretation
pilepu		
piglepu		
kewamo		
kewagomo		
kewagomo		
kewapigonamo		

Figure 14: An example of signal variation for a single target meaning (grey oval) across all three blocks in Mixed + Unshared (rows are ordered on the basis of trial: the top row being the first trial where grey oval was the target and the bottom row being the final trial where grey oval was the target). In the left column is the signal that a speaker produced for the target meaning. The centre column contains the intended meaning of the speaker (which in this case was always the grey oval) and the right column contains what meaning the hearer selected. A correct trial is one where the speaker's intended meaning and the hearer's interpretation are the same meaning. For most of the trials, the hearer selected the incorrect meaning, and the speaker modified the signal in the subsequent trial. Despite this variation pockets of systematic regularity do emerge (e.g., the use of *kewa* to refer to oval).

In our experiment, we did not have any *a priori* assumptions about the relationship between contextual predictability and the emergence of compositionality. This is because lexical and compositional systems are both viable strategies for creating autonomous signals in that a meaning can be inferred out of context. However, as the results demonstrate, compositional systems are overwhelmingly favoured. Why did compositional systems emerge and not lexical ones? One explanation for this preference is that compositional systems reduce hearer uncertainty to a greater extent than lexical systems. From the starting point of our experiment, where the initial training language was ambiguous with respect to colour and shape, the task is for the speaker to successfully convey a set of 16 meanings to the hearer (12 of which are meanings neither the speaker nor the hearer have been previously exposed to). To create a lexical system speakers need to devise 12 signals that not only refer to every new meaning, but are also distinct from the signals they have already used. Hearers then need to infer what each new signal maps onto with relatively few interpretative clues. Assuming hearers are faced with a new meaning in a particular trial, and they only know the signals for the four original meanings, then any given context (array) will contain only one referent they can rule out, with there being 3 possible images to which a signal could map onto (see Figure 15).

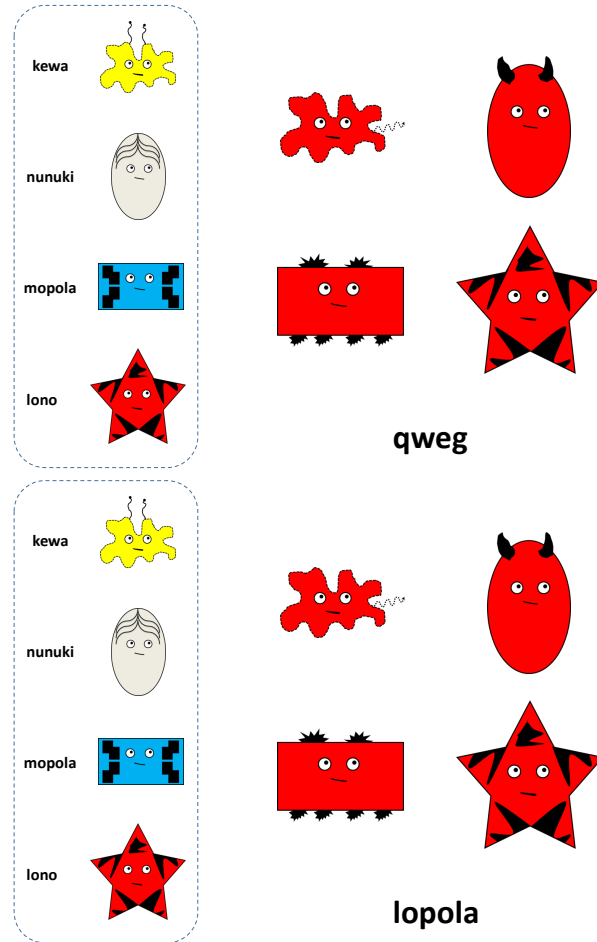


Figure 15: Top: A hypothetical trial where a lexical signal, *qweg*, is invented to convey the intended meaning (which, in this case, is a red rectangle). The box with the dotted line around it contains the signal-meaning pairs both the speaker and hearer share in common. A hearer should, through inference and a bias for mutual exclusivity (Markman & Wachtel, 1988), be able to rule out the red star as the intended meaning (as it already has a conventional signal). But, without any other information, a hearer still has to guess what the signal refers to out of three possible candidates. Bottom: A hypothetical trial where a compositional signal, *lopola*, is invented to convey the intended meaning. Using knowledge that parts of the signal, *pola* and *lo*, belong to the signals for blue rectangle and red star, a hearer should be able to infer that the red rectangle is the intended meaning.

A compositional system minimises inferential effort when contrasted with

a lexical one (Selten & Warglien, 2007). Here, speakers can reuse and modify the signals they were trained on in predictable ways, providing more clues for the hearer to identify the intended meaning. For instance, if the signal *mopola* refers to a blue rectangle and the signal *lono* refers to a red star, then creatively segmenting and reusing components to create a new signal (e.g., *lopola*) helps reduce the burden of hearer inference by narrowing down the space of possible referents in a given context. Imagine the signal *lopola* is used in a Shape-Different context: the rational inference is that *lopola* could only refer to the red rectangle – hearers have knowledge that *pola* is a part of the signal for a blue rectangle and *lo* is a part of the signal for a red star. Given that the context does not have a blue star, and the red star is already designated with signal *lono*, then it is reasonable to infer that the signal *lopola* refers to the red rectangle.

The current experiment was far from exhaustive and can be extended in several ways. One possible extension is to manipulate the number of objects in the referential context (see Rubio-Fernandez, 2016). In terms of identifying the intended meaning, a larger referential context has higher uncertainty than a smaller context, with a hearer needing to sift through more distractors. The reverse is true for the speaker: a larger referential context is more informative than a smaller context. For the largest possible context, where the number of objects is equal to the total number of referents, speakers have access to more information about the necessary distinctions which are globally required by the linguistic system. By contrast, having to discriminate between a single target and distractor only reveals what is locally relevant for discrimination, and is therefore less informative for the speaker in discovering the optimal system for conveying the intended meaning. This tension between what is informative for the speaker versus what is predictable for the hearer is a promising avenue for future research.

The experimental pragmatics literature also offers a few other avenues for stress-testing the relationship between signal autonomy and contextual predictability. For instance, there are various gradations to the possible context-types which remain unexplored, with this set up being restricted to contexts

that unambiguously highlight an informative dimension and background an uninformative dimension. As an example, Frank & Goodman (2012) made a series of manipulations to the context-type, systematically varying whether one, two, or all three of the distractors in a referential context share a feature with the target. Another extension is found in experiments looking at common ground (for reviews, see: Brennan, Galati & Kuhlen, 2010; Konopka & Brown-Schmidt, 2014). In contrast to the current experiment, where the focus is on whether or not the speaker shares access to the hearer’s context, common ground experiments have also investigated the opposite situation: whether or not the hearer shares access to the speaker’s context (e.g., Horton & Keysar, 1996). Incorporating these more fine-grain manipulations to contextual predictability is important if we are to link up predictions about the pragmatic reasoning of participants with differences in signal autonomy.

Lastly, as the communication game had fixed participant roles (they were either a speaker or a hearer), future manipulations could investigate how systems in this experiment vary according to asymmetric and symmetric participant roles (see Moreno et al., 2015 for such a comparison in signalling games). This asymmetric division was necessary in our experiment as we were explicitly interested in the effect of shared information (access to context). Introducing symmetrical participant roles provides the hearer with more knowledge about what distinctions are necessary, and this should decrease the impact of whether the context is shared or unshared. However, having both participants involved in sending and receiving might influence the rate at which interlocutors align on a shared system of communication (e.g., symmetric conditions might take longer to establish a shared system than asymmetric conditions as there are now two participants involved in producing utterances).

There are also important implications for the typological distribution of languages. In this experiment, the final languages varied considerably for conditions where our manipulations to contextual predictability were conflicting (e.g., Shape-Different + Unshared Context and Mixed + Shared Context). If real-world languages are subject to similarly weak constraints, then one general

prediction is that cross-linguistic variation should not straightforwardly reflect a direct relationship between contextual predictability and signal autonomy. Instead, the outcomes will to some extent be historically contingent, albeit with the space of possible languages being bounded by cognitive, contextual, and communicative factors.

5. Conclusion

A good system of communication entails a balance between the demands of the speaker (to reduce their energetic expense) and the demands of the hearer (the speaker needs to provide signals which allow the hearer to identify their intended meaning). We set out to investigate how these two pressures are influenced by the context in which languages were used. By manipulating both Access to Context and Context-Type, we showed that contextual predictability shapes the degree of signal autonomy. When the context is predictable, speakers use this reliable information to reduce their effort in formulating signals, whilst also maintaining the minimal requirement of being informative about the intended meaning. This results in low autonomy: the signals in these systems are dependent on contextual information for disambiguation. However, when the context decreases in predictability, speakers increasingly rely on the signals themselves to reduce uncertainty about the intended meaning, resulting in a greater level of autonomy.

6. References

- Arts, A., Maes, A., Noordman, L., & Jansen, C. (2011). Overspecification facilitates object identification. *Journal of Pragmatics*, 43: 361-374.
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure in mixed-effects models: Keep it maximal. *Journal of Memory and Language*, 68(3): 255-278.
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-

- Effects Models Using lme4. *Journal of Statistical Software*, 67(1): 1-48. doi: 10.18637/jss.v067.i01.
- Bell, A. (1984). Language style as audience design. *Language in Society*, 13: 145-204. doi: 10.1017/S004740450001037X.
- Beuls, K., & Steels, L. (2013). Agent-based models of strategies for the emergence and evolution of grammatical agreement. *PLoS ONE*, 8(3): e58960. doi: 10.1371/journal.pone.0058960.
- Blokpoel, M., et al. (2012). Recipient design in human communication: simple heuristics or perspective taking? *Frontiers in Human Neuroscience*, Special Issue: Towards a neuroscience of social interaction.
- Brennan, S. E., Galati, A., & Kuhlen, A. K. (2010). Two Minds, One Dialog: Coordinating Speaking and Understanding. In B. H. Ross (Ed.), *The Psychology of Learning and Motivation*, 53: 301-344. Burlington: Elsevier.
- Brown, P. M., & Dell, G. S. (1987). Adapting production to comprehension: The explicit mention of instruments. *Cognitive Psychology*, 19: 441-472.
- Channell, J. (1994). *Vague Language*. Oxford: Oxford University Press.
- Clark, H. H. (1996). *Using Language*. Cambridge: Cambridge University Press.
- Clark, H. H., & Carlson, T. B. (1992). Context for Comprehension. In H. H. Clark, *Arenas of Language Use*, 60-77. Chicago: University of Chicago Press.
- Clark, H. H., & Murphy, G. L. (1982). Audience design in meaning and reference. *Advances in Psychology*. 9: 287-299.
- Croft, W. (2000). *Explaining Language Change: An evolutionary approach*. Harlow: Longman.

- Cutting, J., (2007). *Vague Language Explored*. New York, NY: Palgrave Macmillan.
- Dale, R., & Reiter, E. (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19: 233-263.
- Diesendruck, G., & Boom, P. (2003). How specific is the Shape Bias? *Child Development*, 74(1): 168-178.
- Duranti, A. & Goodwin, A. (1992). *Rethinking Context: Language as an interactive phenomenon*. Cambridge: Cambridge University Press.
- Engelkamp, J., Zimmer, H. D. & Mohr, G. (1990). Differential memory effects of concrete nouns and action verbs. *Zeitschrift für Psychologie*, 198: 189-216.
- Evans, V. & Green, M. (2006). *Cognitive Linguistics: An Introduction*. Edinburgh: Edinburgh University Press.
- Fauconnier, G. (1985). *Mental Spaces*. Cambridge, MA: MIT Press.
- Fillmore, C. (1985). Frames and the Semantics of Understanding. *Quaderni Di Semantica*, 6(2): 222-254.
- Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336: 998.
- Franke, M., & Degen, J. (2016). Reasoning in reference games: individual- vs. population-level probabilistic modeling. *PLoS ONE*, 11(5): 1-25.
- Franke, M., & Jäger, G. (2016). Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft*, 35(1): 3-44.

- Freyd, J. (1983). Shareability: The social psychology of epistemology. *Cognitive Science*, 7(3): 191-210.
- Galantucci, B., Garrod, S., & Roberts, G. (2012). Experimental Semiotics. *Language and Linguistics Compass*, 6: 477-493. doi: 10.1002/lnc.
- Gärdenfors, P. (2000). *Conceptual Spaces: the geometry of thought*. Cambridge, MA: MIT Press.
- Gil, D. (2005). Word order without syntactic categories: How Riau Indonesian does it. In A. Carnie, H. Harley, & S. A. Dooley (Eds.), *Verb First: On the Syntax of Verb-initial Languages*: 243-264. Amsterdam: John Benjamins.
- Goffman, E. (1974). *Frame analysis: An essay on the organization of experience*. Cambridge, MA: Harvard University Press.
- Grodner, D. & Sedivy, J. (2011). The effects of speaker-specific information on pragmatic inferences. In N. Pearlmuter & E. Gibson (Eds.). *The processing and acquisition of reference*: 239-272. Cambridge, MA: MIT Press.
- Harrell Jr, F. E., et al. (2014). *Hmisc: Harrell Miscellaneous*. R package version 3.14-5. <http://CRAN.R-project.org/package=Hmisc>.
- Haspelmath, M. (2004). On directionality in language change with particular reference to grammaticalization. In O. Fischer, M. Norde & H. Perridon (Eds.), *Up and down the cline: The nature of grammaticalization (Typological Studies in Language, 59)*: 17-44. Amsterdam: John Benjamins.
- Heller, D., Skovbrotten, K., & Tanenhaus, M. K. (2009). Experimental evidence for speakers' sensitivity to common vs. privileged ground in the production of names. *PRE-CogSci Workshop on the Production of Referring Expressions*. Amsterdam, Netherlands.
- Hopper, P. J., & Traugott, E. C. (2003). *Grammaticalization*. 2nd Edition.

Cambridge: Cambridge University Press.

- Horn, L. R. (1993). Economy and redundancy in a dualistic model of natural language. In S. Shore & M. Vilkuna (Eds.), *SKY 1993: 1993 Yearbook of the Linguistic Association of Finland*: 33-72.
- Horton, W. S., & Keysar, B. (1996). When Do Speakers Take into Account Common Ground? *Cognition*, 59(1): 91-117.
- Hurford, J. R. (2011). *The Origins of Grammar: Language in the Light of Evolution II*. Oxford: Oxford University Press.
- Isaacs, E. A., & Clark, H. H. (1987). References in conversation between experts and novices. *Journal of Experimental Psychology*, 116: 26-37. doi: 10.1037/0096-3445.116.1.26.
- Kay, P. (1977). Language evolution and speech style. In B. G. Blount & M. Sanches (Eds.), *Sociocultural dimensions of language change*: 21-33. New York, NY: Academic Press.
- Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, 11: 323-328.
- Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *PNAS*, 105(31): 10681-10686.
- Konopka, A. E. & Brown-Schmidt, S. (2014). Message encoding. In V. Ferreira, M. Goldrick, and M. Miozzo (Eds.), *The Oxford handbook of language production*: 3-20. New York: Oxford University Press.
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2016). *lmerTest: Tests in Linear Mixed Effects Models*. R package version 2.0-30.

- Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.
- Levinson, S. C. (2000). *Presumptive meanings: the theory of generalised conversational implicature*. Cambridge, MA: MIT Press.
- Lewis, D. (1969). *Convention*. Cambridge, MA: MIT Press.
- Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/Information theory: Speakers choose shorter words in predictive contexts. *Cognition*, 126(2): 313-318.
- Markman, E., & Wachtel, G. (1988). Childrens use of mutual exclusivity to constrain the meanings of words. *Cognitive psychology*, 157: 121-157.
- Minsky, M. A. (1975). A framework for representing knowledge. In M. Winston (Ed.), *The psychology of computer vision*. Boston, MA: MIT Press.
- Mihatsch, W. (2009). Nouns are THINGS: Evidence for a grammatical metaphor? In K. U. Panther, L. L. Thornburg & A. Barcelona. *Metonymy and Metaphor in Grammar*. Amsterdam, Netherlands: John Benjamins.
- Moreno, M., & Baggio, G. (2015). Role Asymmetry and Code Transmission in Signaling Games: An Experimental and Computational Investigation. *Cognitive Science*, 39(5): 918-943. doi: 10.1111/cogs.12191.
- Nadig, A. S. & Sedivy, J. C. (2002). Evidence of perspective-taking constraints in childrens on-line reference resolution. *Psychological Science*, 13: 329-336.
- Nettle, D. (1999). *Linguistic Diversity*. New York, NY: Oxford University Press.
- Nowak, I., & Baggio, G. (2016). The emergence of word order and morphology in compositional languages via multigenerational signaling games. *Journal of Language Evolution*, Advance Access: 1-14. doi: 10.1093/jole/lzw007.

- Olson, D. R. (1970). Language and thought: Aspects of a cognitive theory of semantics. *Psychological Review*, 77: 257-273.
- Parikh, P. (2001). *The Use of Language*. Stanford University: CSLI Publications.
- Partee, B. (1984). Compositionality. In F. Landman and F. Veltman (Eds.), *Varieties of Formal Semantics*: 281-312. Reprinted in B. H. Partee. (2004). *Compositionality in Formal Semantics: Selected Papers by Barbara H. Partee*: 153-181. Oxford, UK: Blackwell Publishing.
- Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, 27: 891-110.
- Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, 122: 280-291.
- Pleyer, M., & Winters, J. (2014). Integrating Cognitive Linguistics and language evolution research. *Theoria et Historia Scientiarum*, 11: 19-43.
- Ramscar, M. & Port, R. (2015). Categorization (without categories). In E. Dawbroska & D. Divjak (Eds.), *Handbook of Cognitive Linguistics*. De Gruyter Mouton.
- Rohde, H., et al. (2012). Communicating with cost-based implicature: A game-theoretic approach to ambiguity. *Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue*.
- Rubio-Fernández, P. (2016). How redundant are redundant colour adjectives? An efficiency-based analysis of color overspecification. *Frontiers in Psychology*. Special issue on Models of Reference.
- Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274(5294): 1926-1928.

- Scott-Phillips, T. C. (2015). *Speaking Our Minds: Why human communication is different, and how language evolved to make it special*. Basingstoke: Palgrave Macmillan.
- Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research*, 32: 3-23.
- Sedivy, J. C. (2005). Evaluating explanations for referential context effects: Evidence for Gricean Mechanisms in Online Language Interpretation. In J. C. Trueswell & M. K. Tanenhaus (Eds.), *Approaches to studying world situated language use: Bridging the language as product and language as action traditions*, Cambridge, MA: MIT Press.
- Selten, R., & Warglien, M. (2007). The emergence of simple languages in an experimental coordination game. *PNAS*, 104: 7361-7366.
- Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, 133: 140-155.
- Silvey, C., Kirby, S., Smith, K. (2015). Word meanings evolve to selectively preserve distinctions on salient dimensions. *Cognitive Science*, 39(1): 212-26.
- Snow, C. E., Barnes, W. S., Chandler, J., Hemphill, L & Goodman, I. F. (1991). *Unfulfilled Expectations: Home and School Influences on Literacy*. Cambridge, MA: Harvard University Press.
- Sperber, D. & Wilson, D. (1995/2005). *Relevance: Communication and Cognition*. 2nd Edition. Oxford/Cambridge: Blackwell Publishers.
- Steels, L. (1999). *The Talking Heads Experiment Volume 1: Words and Meanings*. Brussels: Best of Publishing.

- Steels, L. (2003). Evolving grounded communication for robots. *Trends in Cognitive Sciences*, 7(7): 308-312. doi: 10.1016/S1364-6613(03)00129-3.
- Steels, L. (2012). Self-organization and Selection in Cultural Language Evolution. In L. Steels (Ed.), *Experiments in Cultural Language Evolution*, 1-37. Amsterdam: John Benjamins.
- Tamariz, M. & Kirby, S. (2016). The cultural evolution of language. *Current Opinion in Psychology*, 8: 37-43.
- Terkourafi, M. (2009). On de-limiting context. In A. Bergs and G. Dielwald (Eds.), *Contexts and Constructions*: 17-42. Amsterdam: John Benjamins.
- Tily, H., & Piantadosi, S. T. (2009). Refer efficiently: Use less informative expressions for more predictable meanings. In *Proceedings of the workshop on the production of referring expressions: Bridging the gap between computational and empirical approaches to reference*.
- Winters, J., Kirby, S., & Smith, K. (2015). Languages adapt to their contextual niche. *Language and Cognition*, 7(3): 415-449.
- Wittgenstein, L. (1953). *Philosophical Investigations*. G.E.M. Anscombe and R. Rhees (Eds.), Trans. G.E.M. Anscombe. Oxford: Blackwell.
- Wray, A., & Grace, G. W. (2007). The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua*, 117(3): 543-578. doi: 10.1016/j.lingua.2005.05.005.
- Zhang, Q. (2011). Elasticity of Vague Language. *Intercultural Pragmatics*, 8(4): 571-599.
- Zipf, G. (1949). *Human behavior and the principle of least effort*. New York, NY: Addison-Wesley.

4.4 Conclusion

Establishing common ground in communication requires speakers and hearers negotiate a conventional system of form-meaning mappings. These mappings emerge incrementally, resulting from interactions between context (*discrimination pressure*), cognition (*generalisation pressure*), and communication (*coordination pressure*). Here, we introduced strong discrimination and generalisation pressures, and investigated how solutions to the coordination pressure are influenced by manipulations to Access to Context and Context-Type. Together, these two manipulations correspond to contextual predictability, with the experiment in this chapter specifically testing the hypothesis that contextual predictability is causally related the degree of signal autonomy.

When the context is predictable, speakers use this reliable information to reduce their effort in formulating signals, whilst also maintaining the minimal requirement of being informative about the intended meaning. This results in low autonomy: the signals in these systems are dependent on contextual information for disambiguation. However, when the context decreases in predictability, speakers become increasingly reliant on the signals themselves to reduce uncertainty about the intended meaning, resulting in a greater level of autonomy. Lastly, we showed how conditions where our manipulations to contextual predictability are conflicting (Shape-Different + Unshared Context and Mixed + Shared Context) produce more heterogeneous outcomes than conditions where contextual predictability is reinforcing (Shape-Different + Shared Context and Mixed + Unshared Context).

Taken together, these results show that context, cognition, and communication all interact to shape the organisation and structure of language.

Chapter 5

Linguistic diversity and traffic accidents

5.1 Introduction

Experiments are not the only way one might go about investigating the relationship between context and language structure. The Kantian philosopher, Wilhelm Windelbrand (1894/1998), described two distinct methodological traditions (Hurlburt & Knapp, 2006). The first of these, termed *idiographic*, refers to a narrow focus on specific phenomena, often employing case studies, unstructured observation, and other qualitative methods as a means of discovery. The notion of context has been a guiding principle for using qualitative methods to study language: as no two contexts are the same, language use is subject to considerable variation, with research efforts being focused on providing rich, detailed descriptions (e.g., see *ethnography of communication*: Hymes, 1987).

By contrast, *nomothetic* approaches make use of large-scale surveys, experiments, and statistical analyses in an effort to seek general, law-like explanations. In particular, the last decade or so has seen a rise in the number of studies taking advantage of large-scale, cross-cultural datasets, and newly available statistical techniques, to investigate the relationship between language structure and non-linguistic variables (Dediu & Ladd, 2007; Hay & Bauer, 2007; Lupyan & Dale, 2010; Atkinson, 2011; Chen, 2013; Roberts, Winters & Chen, 2015; Everett, Blasi & Roberts, 2015; Lewis & Frank, 2016).

Lupyan & Dale (2010) provide an illustrative example for how one might go about investigating the relationship between language structure and context. Here, the authors use three demographic variables – population size, geographic spread, and degree of language contact – as a proxy for social context. The

idea being that languages with small populations, occupying a geographically restricted area and having relatively few linguistic neighbours, closely corresponds to situations which call for *esoteric*, or intra-group, communication: languages predominately used in societies comprising of close intimates, where the social structure results in individuals sharing a high degree of contextual and cultural knowledge. This is in contrast to *exoteric*, or inter-group, communication where the social structure leads to a greater proportion of non-native adult learners and individuals who share less common cultural and contextual knowledge with one another. Differences between esoteric and exoteric forms of communication are therefore predicted to shape the structure of language as it is learned and used (also see: Trudgill, 2004; Wray & Grace, 2007; Trudgill, 2011; Hurford, 2011; Nettle, 2012).

Operationalising social context in this way allowed Lupyan & Dale (2010) to indirectly test whether languages used for esoteric communication have greater levels of morphological complexity¹ than languages used for exoteric communication. Indeed, this is what they found in a sample of over 2000 languages, with population size being the strongest demographic predictor of morphological complexity: small populations are more likely than large populations to use languages which rely on morphological strategies to encode semantic distinctions. Still, even though population size is a significant predictor of morphological complexity, all this establishes is that a pattern exists between two variables. Lupyan & Dale have a preferred explanation for why morphological complexity persists in esoteric communities and undergoes simplification in exoteric communities². But statistical analyses are not a causal explanation of such relationships – and they do not necessarily rule out competing explanations for the fit between language structure and social context (see: Dale & Lupyan, 2012; Nettle, 2012; Atkinson, Kirby & Smith, 2016).

This chapter highlights several problems with using a statistical approach. First, spurious correlations are a common property of cross-cultural datasets – historical descent, geographic diffusion, and high signal-to-noise ratios all play a role in shaping the relationship between variables. Controlling for these factors is contingent not just on the sophistication of the statistical methods, but also the

¹Morphological complexity is defined as the use of morphological strategies over lexical ones to encode semantic distinctions like evidentiality, future tense, and epistemic possibility (Lupyan & Dale, 2010).

²Lupyan & Dale (2010) argue that morphological paradigms are difficult for adults to learn; therefore, languages exoteric communities are more likely to lose complex morphological systems. They also provide a tentative explanation for why complex morphology is maintained in esoteric communities: morphological overspecification facilitates child learning by providing multiple cues during language acquisition.

quality of the data being used to perform the correlation. Second, in line with Nettle (2007), I argue statistical studies are better viewed as *hypothesis-generating* as opposed to *hypothesis-testing*. This is especially relevant when there are no *a-priori* assumptions about the directionality of the effect. Lastly, the chapter concludes by stating how these approaches are not suitable for investigating the causal relationship between language structure and context, and instead argues for the use of laboratory experiments.

5.2 Author contributions

The following section contains a paper which was co-authored with Sean Roberts and published in PLoS One. Each author jointly contributed to the writing of the paper and the analyses contained within.

5.3 Roberts & Winters (2013): Linguistic diversity and traffic accidents

Linguistic Diversity and Traffic Accidents: Lessons from Statistical Studies of Cultural Traits

Seán Roberts^{1*}, James Winters²

1 Seán Roberts Max Plank Institute for Psycholinguistics, Nijmegen, The Netherlands, **2** James Winters Language Evolution and Computation Research Unit, School of Philosophy, Psychology and Language Sciences, The University of Edinburgh, Edinburgh, United Kingdom

Abstract

The recent proliferation of digital databases of cultural and linguistic data, together with new statistical techniques becoming available has led to a rise in so-called *nomothetic* studies [1–8]. These seek relationships between demographic variables and cultural traits from large, cross-cultural datasets. The insights from these studies are important for understanding how cultural traits evolve. While these studies are fascinating and are good at generating testable hypotheses, they may underestimate the probability of finding spurious correlations between cultural traits. Here we show that this kind of approach can find links between such unlikely cultural traits as traffic accidents, levels of extra-marital sex, political collectivism and linguistic diversity. This suggests that spurious correlations, due to historical descent, geographic diffusion or increased noise-to-signal ratios in large datasets, are much more likely than some studies admit. We suggest some criteria for the evaluation of nomothetic studies and some practical solutions to the problems. Since some of these studies are receiving media attention without a widespread understanding of the complexities of the issue, there is a risk that poorly controlled studies could affect policy. We hope to contribute towards a general skepticism for correlational studies by demonstrating the ease of finding apparently rigorous correlations between cultural traits. Despite this, we see well-controlled nomothetic studies as useful tools for the development of theories.

Citation: Roberts S, Winters J (2013) Linguistic Diversity and Traffic Accidents: Lessons from Statistical Studies of Cultural Traits. PLoS ONE 8(8): e70902. doi:10.1371/journal.pone.0070902

Editor: Frank Emmert-Streib, Queen's University Belfast, United Kingdom

Received: January 23, 2013; **Accepted:** June 24, 2013; **Published:** August 14, 2013

Copyright: © 2013 Roberts, Winters. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: Seán Roberts was partly supported by an ESRC grant ES/G010277/1. James Winters is supported by an AHRC grant AH/K503010/1. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: sean.roberts@mpi.nl

Introduction

Recent studies have been uncovering some surprising links between cultural traits. For example, between chocolate consumption and the number of Nobel laureates a country produces [9], between the number of phonemes in a language and distance from East Africa [3], between a language's tense system and the propensity to save money [2], between the quality of the sounds of a language with the amount of extra-marital sex [6] and genetic influences on political outlooks [4,5].

Nomothetic studies (statistical analyses of large-scale, cross-cultural data) are possible because of recently available, large-scale databases and new statistical techniques which give social scientists more statistical power to explore the relationships between cultural phenomena. They are quick and easy to perform. However, there are several potential problems with this type of study. While it is common knowledge that correlation does not imply causation, there are few studies that utilise methods to address the problems caused by cultures being related by descent (Galton's problem, [10], see [11]) and by geographic diffusion [12]. Furthermore, the data used in these studies is inherently coarse, which can create apparent correlations. There is also the problem of inverse sample-size: with larger amounts of data, a spurious correlation becomes more likely.

These problems combine to increase the likelihood of finding correlations between cultural traits. In this paper we demonstrate that it is possible to link a wide variety of cultural traits in a chain

of correlations, all of which may seem rigorous, but some of which are not plausibly causal. In fact it may be possible to find apparently rigorous evidence for any hypothesis. It is also tempting to fit post-hoc hypotheses to correlations that fall out of nomothetic studies. However, without a proper awareness of the problems, this kind of study could be damaging to the direction of research and public policy.

The inter-connectedness of cultural traits that we demonstrate raises problems for the usefulness of statistical analyses as independent sources of knowledge. However, we suggest that nomothetic studies should be seen as hypothesis-generating tools that can work with and direct other methods such as idiographic studies, computational modelling, experiments and theoretical work [13,14]. We also suggest some methods that might improve statistical inference and insight in nomothetic studies, including phylogenetic techniques and inferred causal graphs [15]. To our knowledge, this is the first application of high-dimensional causal graph inference to cultural and linguistic data.

The paper is organised in the following way. First, we summarise some nomothetic case-studies. We outline some problems facing nomothetic studies and suggests some criteria for evaluating them. Our results section demonstrates a chain of statistically significant links between cultural traits, followed by a short discussion. Finally, we suggest some solutions to the problems discussed.

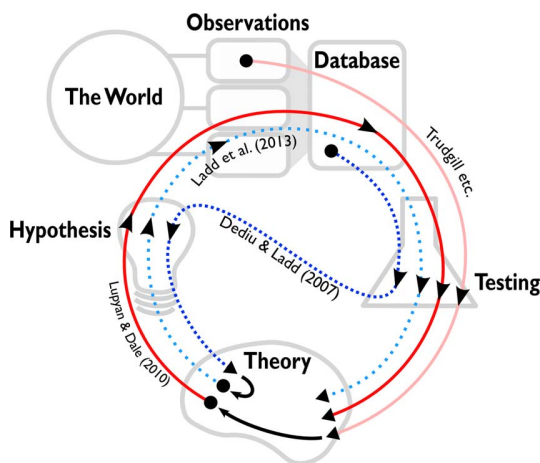


Figure 1. The scientific processes of different nomothetic studies. Observations are drawn from the world, either as idiographic studies or experiments. These observations can be compiled into large-scale cross-cultural databases. Scientific elements include theory, hypotheses and testing. Trajectories indicate the process of different studies. Processes start at a dot and continue in the direction indicated by the arrows. The ideal trajectory is the following: A theory generates a hypothesis. The hypothesis suggests data to collect, which is then tested. The results of the test feed back into the theory. Lupyan & Dale (2010) follow this trajectory, although they take their data from a large-scale cross-cultural database. Lupyan & Dale's theory was generated by previous testing of (small-scale) observations by Trudgill and others. The trajectory of Dediu & Ladd's study differs in two ways. First, the trajectory starts with large-scale cross-cultural data rather than small-scale observations. Secondly, the testing generates the hypothesis, which suggests a theory. However, Ladd et al. (2013) use this theory to motivate a hypothesis which is tested on experimental data. Since developing theories from small-scale observations takes time and effort, Dediu & Ladd's study has effectively jump-started the conventional scientific process.
doi:10.1371/journal.pone.0070902.g001

Nomothetic Studies

One example of a nomothetic study used the World Atlas of Language Structures (WALS) [16] database to demonstrate that a community's size is related to the morphological complexity of its language [1]. This is a well controlled statistical test which is robust across language families. The suggested mechanism behind this link, motivated by prior theory (e.g. [17–19]), is the difference between adult and child language learning. Because larger communities are more likely to have more adult second language learners, and adults are worse at learning morphology than children [20], this puts pressure on languages of large communities to become less morphologically complex over time.

Another study discovered a link between areas with a prevalence of a recently mutated geneotype and populations with tonal languages (languages where lexical contrasts can be made by altering pitch patterns) [7]. This generated the hypothesis that linguistic structure could be affected by small genetic biases over time. Because the baseline level of chance correlation is difficult to estimate, the statistical significance was computed by comparing the strength of the link to the strength of the link between thousands of other linguistic and genetic variables. By demonstrating that the hypothesised link was stronger than competing hypotheses, a convincing claim was made for the further experimental investigation of this hypothesis. In order to develop the basis of the general theory, a follow-up experimental study found support for part of the hypothesis in that there are individual

differences in the perception of pitch [21], and a computer simulation demonstrated that such differences could influence linguistic structure in the long-term [22].

A number of studies have demonstrated links between a community's size and the number of contrasting sounds (phonemes) in its language. Hay & Bauer demonstrate a positive correlation between population size and phoneme inventory size [8] (replicated in [3,23]). However, recent analyses using larger samples and accounting for the relatedness of languages find no such correlation [24,25]. While the original results might be debatable, and despite the proposed link between phoneme inventories and social structure being well-established (e.g. [26]), the debate surrounding the original nomothetic study did offer the opportunity for the development and application of a wide variety of statistical techniques. This includes the use mixed effects models that can control for nested data by placing predictors at differing levels [24,27].

Ecological aspects have also been shown to predict linguistic variables. Correlations are reported in [28] between the average sonority of a language – the average amplitude of its phonemes – and the local climate and ecology. The proposed hypothesis includes people in warmer climates spending more time outdoors, and sonorous sounds being more effective at communicating at a distance. This finding was extended to account for cultural features such as the amount of baby-holding, levels of literacy and attitudes towards sexual promiscuity [6]. The link with sexual attitudes is hypothesised as being due to sexual inhibition discouraging speaking with a wide open mouth. Below, we show that population size also correlates with these variables.

Nomothetic studies can also straddle relatively disparate fields. For example, two studies find a correlation between the distribution of political attitudes (individualist versus collectivist) and the prevalence of a gene involved in the central neurotransmitter system 5-HTTLPR [4,5]. The social sensitivity hypothesis suggests that, because alleles of this gene affect the likelihood of a depressive episode under stress [29], communities with a higher prevalence of this gene will require more social support. Therefore, these communities will develop to be more collectivist rather than individualist. However, a missing element of these alleles emerged in the first place. By exploring the inter-connectedness of many different variables, we develop a hypothesis which suggests that migration and environmental conditions could bring about this distribution (see the section 'Causal graphs').

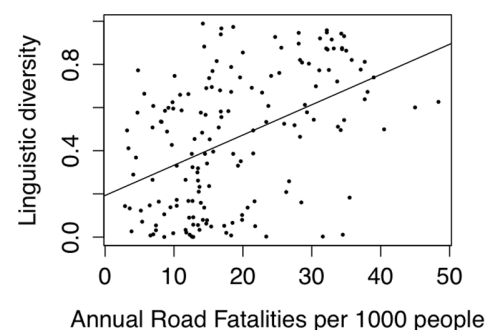


Figure 2. Linguistic diversity and traffic accidents. Countries' linguistic diversity (Greenberg diversity index) as a function of the annual traffic fatalities per 1000 people, with linear regression line.
doi:10.1371/journal.pone.0070902.g002

Media attention

Nomothetic studies often demonstrate surprising links between cultural phenomena. For this reason, they often receive media attention. For example, based on research that flavonoids found in chocolate benefit cognitive function [30], a study demonstrated that countries which have a higher per-capita consumption of chocolate produce more Nobel laureates [9]. The study used a simple linear correlation, without controlling for any other factors, yet received a large amount of media attention [31–33]. Even though the study may have intended the correlation to be interpreted as an example of spuriousness, it failed to control for other factors and possible confounds. This is an example of a misapplication of statistical techniques. While this particular study may seem harmless, below we use the same data to demonstrate correlations which appear to have more serious implications for public attitudes and policy.

Another study that has received media attention is the finding that speaking a language that has an overt morphological future tense predicts economic behaviour such as the propensity to save money [2]. This study was discussed before publication in public forums online [34–36] and in the media [37–41]. The media typically exaggerate the implications of this type of finding and try to link it to current events rather than emphasise the long-term change implied in most studies. For instance, one popular science review of study [2] suggested “Want to end the various global debt crises? Try abandoning English, Greek, and Italian in favor of German, Finnish, and Korean.” [38].

Problems

In this section we review three problems that cause spurious correlations in nomothetic analyses of cultural phenomena.

Galton’s problem

One of the better-known issues facing nomothetic researchers is that of *Galton’s Problem* [10]. Named after Sir Francis Galton, following his observation that similarities between cultures are also the product of borrowing and common descent, Galton’s Problem highlights that researchers must control for diffusional and historical associations so as to not inflate the degrees of freedom in a sample [42]. For example, the likely magnitude of a correlation emerging between two independent traits is much higher if the traits diffuse geographically than if they change randomly [34]. Cultural traits, then, form a complex adaptive system [43] where some links are causal and some links are accidents of descent. For this reason, we would expect to see spurious correlations appearing between unlikely cultural variables.

Ascertaining the degree of independence between cases is a concern that has a long history in cross-cultural research [44]. Numerous methods have been proffered as potential solutions, notably: spatial autocorrelation, phylogenetics, generalised linear mixed models [12]. One debated difference is the amount of horizontal transmission that occurs in cultural traits [45–48]. While there are well-developed models for genetic evolutionary change that are used in phylogenetic analyses [49], it is less clear whether they are suitable for assessing cultural change. Complicating this is the difficulty of identifying cultural traits in the past due to a lack of comparative evidence and the transience of cultural traits such as spoken language.

Large datasets and complex relationships are dealt with regularly in fields like genetics. However, there is an active debate about the role of statistics in causal inference [50]. Neuroscience studies involving brain imaging also deal with large, complex datasets. However, spurious correlations are also a problem here

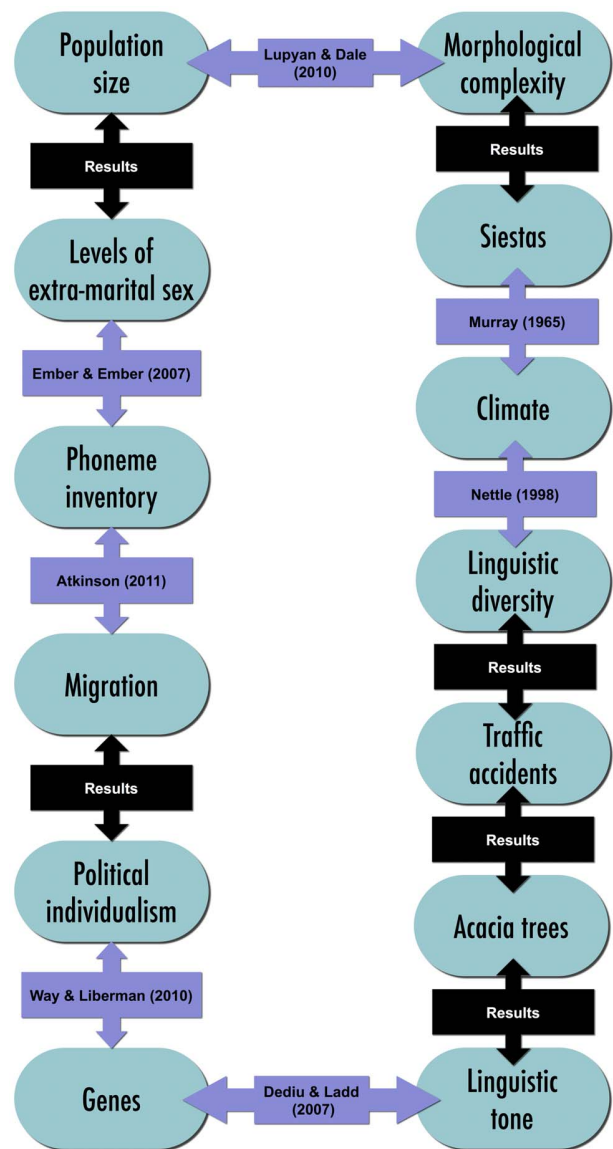


Figure 3. Chains of spurious correlations. Statistical links can be found between these cultural traits. Links from previous studies are labelled with the authors’ names. The links from the results section of the current study are labelled ‘results’. doi:10.1371/journal.pone.0070902.g003

[51,52], and the inference based on some advanced techniques have been recently questioned [53]. Despite an awareness of the problem, there are few studies with a sophisticated approach to addressing it. In general, review of statistics used in studies of culture and language may be less rigorous than in other fields [54]. This might suggest that, for researchers, the crux of the problem is a lack of tools, not a lack of awareness of the problem.

Distance from data: Are linguists the main drivers of changes in consonant inventory sizes?

Nomothetic studies often use databases that exhibit a distance from the real data. This is particularly salient when the datasets consist of statistically rare observations i.e. one researcher generated all the data for one particular data point. The amount of variance and selection bias introduced via the process of getting

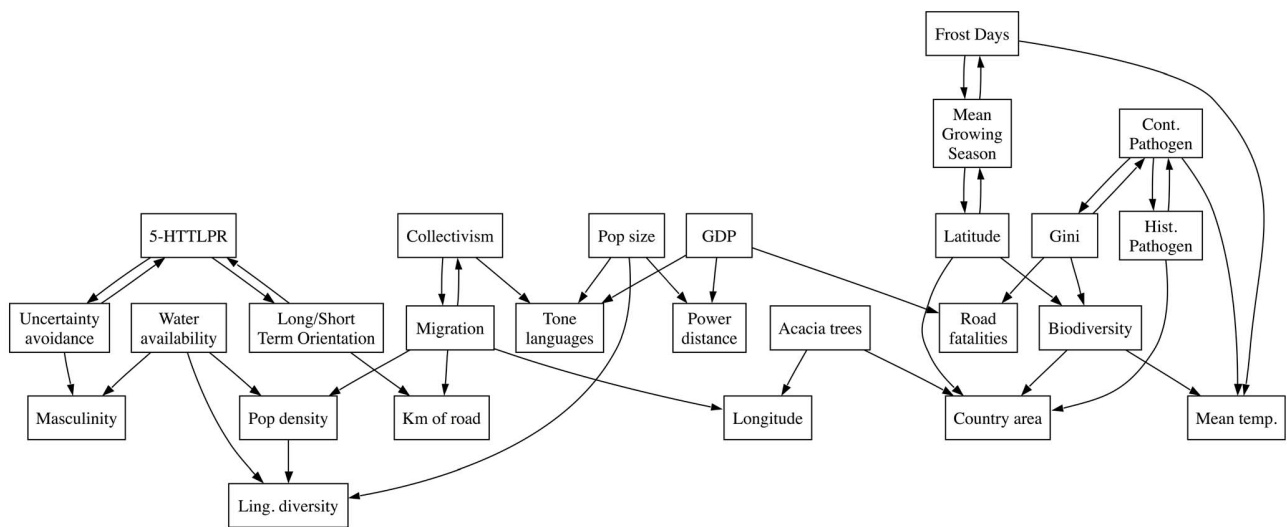


Figure 4. The most likely directed, acyclic graph of causal relationships between different variables in this study. Boxes represent variables and arrows represent suggested causal links going from a cause to an effect. See the methods section for details. doi:10.1371/journal.pone.0070902.g004

from actual data-collection to the database in question can be problematic in terms of analysis.

An illustrative example is found in classifying the size of a particular constant inventory for WALS. WALS determines its consonant inventory size data by binning raw consonant counts into the following categories: small (6–14), moderately small (15–18), average (22 ± 3), moderately large (26–33) and large (34 or more) [55]. These categorical variables are useful for the context in which WALS was created: to highlight the geographic distribution of typological diversity around the globe. However, such coarse lumping into categorical variables might inflate error, especially when the variables could lend more weight to finding a particular correlation than another.

Still, there can also be considerable distance between the observations of different field linguists. Take the reports of consonant inventory sizes for the Wichí language – a member of the Matacoan language family spoken in various parts of South America’s Chaco region [56]. For instance, in 1981 when Antonio Tovar published an article on the Wichí’s phoneme inventory [57], he arrived at a figure of 22 consonants. Jump forward 13-years to 1994 and Kenneth Claesson’s report [58] would tell you the Wichí are down to just 16 consonants. This is just one of what is likely be many examples of huge degrees of variation in linguistic observations for rare languages. The difference in reports would be enough to change the categorical value in WALS from an average consonant inventory size to a moderately small one.

There are several explanations for the variance in such reports. Some instances could be genuine differences between speech communities in the form of dialectal variation. Other reasons take the form of theoretical motivation. Claesson, for instance, chose to omit glottalized consonants from his description of Wichí. His rationale being that these “are actually consonant clusters of a stop followed by a glottal stop” [56,37–38]. In summary, both sources of data are sensitive to the biases of the researchers: for each language, or dialect, these observations are reliant on the choices of potentially one researcher, at a very specific point in time, and with only a finite amount of resources. We believe such sources of variance are not limited to phoneme inventory data, but rather are endemic in these sorts of data, which leads to the problem of having “too many variables (but too little data per variable)” [59].

Inverse sample size problem

Whilst we believe big data is a valuable resource for social scientists, the type of big data collected, as well as the types of questions asked in relation to these datasets, are of a fundamentally different nature to those found in other areas that rely on large datasets. Pick up any statistical textbook and it is likely you will read something along the lines of “as is intuitively obvious, increases in sample size increase statistical power” [60]. This is certainly true on an absolute basis where there is a decrease in the noise-to-signal ratio. For instance, the extremely small sample sizes in neuroscience are probably responsible for the overestimates of effect size and low reproducibility of results ([52]; but also see [61] for a more general discussion on this problem across all sciences). We have also seen great successes in physics where large amounts of data were crucial in the discovery of the Higgs Boson [62,63] or in astronomy with the spectroscopic survey of millions of stars (the Sloan digital sky survey [64]). Yet, as Gary Marcus recently noted, large datasets in physics are characterised by certain properties:

“Big Data can be especially helpful in systems that are consistent over time, with straightforward and well-

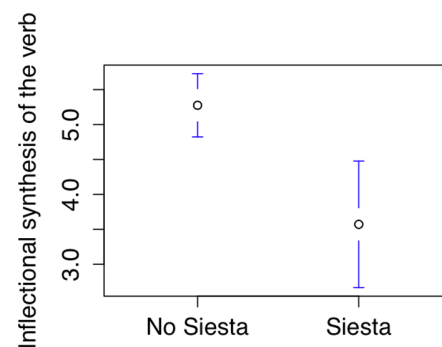


Figure 5. Relationship between siestas and number of grammatical categories a verb can take. doi:10.1371/journal.pone.0070902.g005

characterised properties, little unpredictable variation, and relatively little underlying complexity” [65].

It is tempting to apply the same principled reasoning to the nomothetic approach to culture. However, nomothetic studies tend to rely on data that falls on the opposite end of the spectrum: these datasets tend to be incomplete, complex and based on inconsistent criteria. Problems such as those in the case of the Wichí’s consonant inventory are just some issues that we know about. There are likely to be unknowable confounds that increase the amount of hidden error in a particular sample. As such, the types of data found in nomothetic approaches are faced with an inverse sample size problem: the noise-to-signal ratio increases exponentially with an increase in the size of the dataset. This is not to say that small data has a higher signal-to-noise ratio. But it does raise the problem that these various confounding factors in large datasets make finding a signal in amongst the noise increasingly difficult. As Taleb cogently puts it:

“This is the tragedy of big data: The more variables, the more correlations that can show significance. Falsity also grows faster than information; it is nonlinear (convex) with respect to data” [59].

Evaluating Nomothetic Studies

We can use two of the issues above to evaluate nomothetic studies. First, the extent to which the experimental hypothesis is embedded in an existing theoretical framework. This relates to the hypothesised mechanism that causes the correlation that is presented. The second issue is the extent to which the study attempts to control for alternative hypotheses, particularly involving the historical relatedness of the observations. This relates to the strength of the correlation.

The interaction between these two issues lead to four types of study. First, there are studies that are motivated by prior theoretical and experimental work and are statistically rigorous. For example, the relationship between population size and morphological complexity (see above). This type of study can be valuable for testing hypotheses, generating hypotheses and acting as a catalyst for interdisciplinary work [14].

The second type of nomothetic study, which may also be valuable, includes studies which may not have been motivated by prior theories, but rigorously demonstrate that the hypothesised link is statistically sound. For example, Dediu and Ladd’s study of

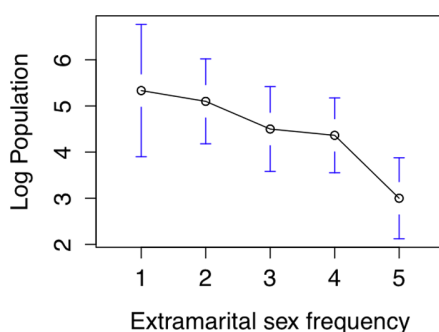


Figure 6. Relationship between population size and frequency of extramarital sex in a society.
doi:10.1371/journal.pone.0070902.g006

genetic correlates of speakers of tonal languages demonstrated that their hypothesised link was significantly stronger than thousands of similar links. This type of study can be very useful for discovering new links that can motivate new avenues of research [13], especially when direct evidence is difficult to obtain. The link between tone and specific genes might have taken much longer to discover by small-scale studies. However, the statistical analysis does not directly support the hypothesised mechanism behind the link [14]. This must be done with methods other than nomothetic studies, such as experiments (e.g. [21]).

It is not always easy to judge whether the right controls are in place. Below we demonstrate a correlation between traffic accidents and linguistic diversity. This was not motivated by a prior theory, but it remains robust against controls for many factors. This type of study can be difficult to evaluate because the factors may be related in complex ways that are difficult to intuit about, or simply that the probability of a spurious correlation is increased in studies with large datasets.

The other two types of study may be detrimental. Those that are grounded in existing theories, but are poorly controlled risk missing hidden complexities which might challenge or develop the theory. For example, the study linking chocolate consumption and Nobel laureates (see above) was based on experimental findings on the cognitive benefits of chocolate. However, the statistical method was simply a linear correlation without any control variables. We find that the correlation does not remain significant when controlling for gross domestic product (GDP) and climate (see methods). More importantly, it is difficult to see what extra insight the this study provides over the controlled experiments that motivated it. This particular study has certainly gained public attention, but this might be dangerous if public opinion or policy is affected by poorly controlled studies.

Finally, studies that are not grounded in theory and are also poorly controlled can be misleading. It is difficult to distinguish these studies from ‘fishing’ for correlations from a large set of variables, then fitting a post hoc hypothesis to the strongest outcomes. As we demonstrate below, since cultural phenomena are subject to non-intuitive constraints, such as Galton’s problem, it is relatively easy to produce evidence for a link between almost any two cultural variables that has the appearance of rigour. For example, we find that the per-capita consumption of chocolate also predicts the number of serial killers and rampage killers a

Table 1. Population size and extramarital sex.

	Coefficient	Std. Error	t value	Pr(> t)
(Intercept)	7.40	1.94	3.82	0.01 *
Population density	0.32	0.11	2.85	0.01 *
Premarital sex frequency	-0.39	0.21	-1.90	0.07
Premarital sex deterrence	-0.22	0.34	-0.65	0.52
Extramarital sex frequency	-0.38	0.17	-2.21	0.04
Extramarital sex deterrence	-0.31	0.30	-1.05	0.31

Results of a regression using population size as the dependent variable and independent variables including population density and four measures of patterns of and attitudes to extramarital sex.
doi:10.1371/journal.pone.0070902.t001

country produces (see methods). There was no prior reason to think that this relationship would hold, apart from the likelihood of cultural traits being correlated. Despite this, it appears to support negative effects of chocolate, in opposition to the positive associations of the study above [9]. There is a danger that these methods could be exploited by researchers, politicians or the media to support particular agendas.

An example from economics highlights this danger. A well-cited study found a correlation between countries with a high ratio of national debt to GDP and countries with slow GDP growth [66]. The authors interpret this as economic growth being stifled by high debt. Although this goes against established theories [67], this interpretation has been widely cited in the media [68] and has been used in testimony before the US senate budget committee in order to support budget cuts [69]. However, the results have recently been shown to be an effect of poor statistical controls and the accidental exclusion of a cluster of related countries [68]. A more careful analysis revealed that countries with high debt to GDP ratios actually had positive growth [68]. Despite this radical change in implication, some commenters are already predicting that it will have little effect on policy, since the statistic was being used opportunistically to support claims for which theoretical arguments were more valid [70]. In this sense, correlational studies can be used as rhetorical devices with the appearance of rigour, but which actually have low explanatory power. Furthermore, damage caused by misleading studies may not be easy to fix.

The potential negative implications of nomothetic approaches can be addressed by applying more rigorous standards to statistical methods and increasing the awareness amongst researchers and the general public of the fragility of simple correlational studies. We hope to contribute to this awareness by demonstrating a chain of surprising links.

Processes

Another way to think about the differences between nomothetic studies is by tracking the way they develop. The two useful types of study follow different processes (see figure 1). The ideal process of a study is for a theory to generate an experimental hypothesis, the hypothesis to suggest data to collect and a way to analyse or test them, and then the results of the analysis to feed back into a better understanding of the theory. The study on the relationship

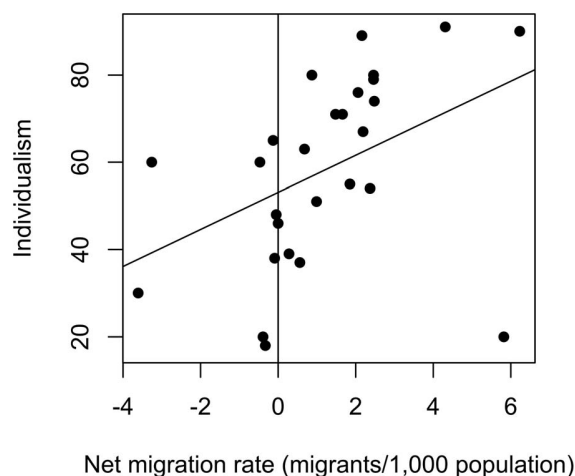


Figure 7. Individualism-Collectivism as a function of migration. Data shown for 28 countries with linear regression line. Large values on the y-axis indicate greater individualism. doi:10.1371/journal.pone.0070902.g007

between population size and morphological complexity [1] follows this process trajectory, although it uses large-scale cross-cultural data. Of course, all theories have to start somewhere, and the theory that the study was based on was developed from small-scale idiographic data. This is an example of how a nomothetic approach can use large-scale data to test hypotheses suggested by small-scale studies.

The study of genetic correlates of linguistic tone [7] had a different trajectory. Here, there was no prior theory. Instead, a pattern in large-scale data suggested a hypothesis which was developed into a theory (see [71]). However, this theory went on to suggest an experimental hypothesis which was tested on small-scale experimental data [21]. This is an example of how a nomothetic study can use large-scale data to generate hypotheses that motivate small-scale, experimental studies.

The two approaches follow different approaches to science. The former fits with a hypothetico-deductivist approach, the latter fits with a more inductive approach to science (although the division between the two approaches is not always clear-cut) [72]. However, the small-scale study in the latter example also followed the more conventional scientific process. In this sense, since developing theories from small-scale observations takes time and effort, the latter nomothetic study jump-started the conventional scientific process.

Results

Chain of correlations

If cultural traits are co-inherited, by descent or horizontal transmission, we should expect to find correlations between many cultural and demographic traits. For instance, the linguistic diversity of a country is correlated with the number of fatalities due to traffic accidents in that country, even controlling for country nominal GDP, per-capita GDP, population size, population density, length of road network, levels of migration, whether the country is inside or outside of Africa (a strong predictor of road fatalities), distance from the equator and absolute longitude ($r = 0.45$, $F(97,10) = 2.03$, $p = 0.003$, see figure 2 and methods). This result is also robust to controlling for the geographic relationships between countries ($r = 0.22$, $p = 0.000001$, see methods).

Table 2. Genes, collectivism and migration.

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	119.0229	30.6819	3.88	0.0009
5-HTTLPR short-short allele prevalence	-1.1268	0.2478	-4.55	0.0002 *
Gini coefficient	-0.5244	0.2977	-1.76	0.0935
Contemporary pathogen prevalence	0.3800	0.8025	0.47	0.6409
Historical pathogen prevalence	-7.7133	6.3512	-1.21	0.2387
GDP	-0.0002	0.0003	-0.71	0.4855
Net migration	5.0025	1.5565	3.21	0.0044 *

A linear regression with levels of collectivism in a country as the dependent variable and independent variables including the prevalence of an allele of the serotonin transporter functional polymorphism 5-HTTLPR, Gross Domestic Product, Gini coefficient, measures of pathogen prevalence and contemporary migration levels.

doi:10.1371/journal.pone.0070902.t002

Furthermore, it is possible to demonstrate a chain of relationships between cultural variables (see figure 3): Linguistic diversity is linked with climate [73]. Climate affects the likelihood of cultural siestas [74]. Cultures that take siestas tend to have languages with less morphological complexity ($t = 3.47$, $p = 0.001$, see methods). Morphological complexity is linked with group size [1]. Group size is linked to the levels of extra-marital sex in a community ($r = -0.54$, $p = 0.001$, see methods). Levels of extra-marital sex have been linked to a language's phoneme inventory [6]. Phoneme inventories have been linked to patterns of migration [3]. Migration patterns are linked to the level of political collectivism in a culture ($r = 0.42$, $p = 0.004$, see methods). Collectivism is predicted by genetic factors [4,5]. There are also genetic correlates of linguistic tone [7]. Tonal languages co-occur with acacia trees ($t = 3.77$, $p = 0.0002$, see methods). To bring the chain full-circle, the presence of *Acacia nilotica* also predicts a greater number of traffic accident fatalities, controlling for linguistic diversity, length of road network, GDP, distance from the equator, population size and population density ($t = 3.26$, $p = 0.0014$, see methods).

Discussion

In the analyses above, we demonstrated a chain of correlations between cultural and demographic features. Some links are well motivated by prior hypotheses and statistically sound (e.g. [1], as discussed above). Others might not have had prior motivation, but are statistically sound and, in some cases, have gone on to be tested by experiments (e.g. [7], as discussed above).

In contrast, some of the studies fail the evaluation criteria discussed in the previous section. Some of the analyses are poorly controlled. For example, the link between acacia trees and tonal languages does not account for obvious environmental features such as temperature and altitude. However, some of the analyses appear statistically sound, but have no prior motivation and are not plausibly causally linked. For example, the link between traffic

accidents and linguistic diversity controlled for many relevant factors. One could hypothesise that miscommunication between speakers of different languages could cause accidents, but it is more likely that a third variable such as the stability of the state explains both linguistic diversity and traffic safety. In this example, the confound is fairly obvious. However, as the number of variables involved increases, and the processes become more complex, it can become increasingly difficult to have intuitions that would lead to this resolution. Political stability might be an obvious control to include for a political scientist, but might not occur to a linguist. The kinds of aspects that nomothetic studies are being used for are typically on the border between two or more disciplines (Genetics and Linguistics [7]; Economics and linguistics [2]; Morphology, language change and demography [1]). Without a broad knowledge of these disciplines, or collaboration, this is exactly the kind of situation which might be difficult to intuit about.

The opposite problem – of knowing which variables to exclude from an analysis – may be equally difficult to answer. Since there is a chance that any cultural traits will be correlated, and since we actually demonstrate some above, there is an argument for including more control variables. For example, if a study investigates linguistic diversity, should it take the number of traffic accidents into account? Worse, since we demonstrate a chain of links, should a study of any of them control for all of the others? That is, if a study is interested in morphological complexity, should it take the collectivism of its speakers into account? For many methods, including more variables reduces statistical power and complicates the analysis. While intuition and theory play a role in knowing what to control for, in the next section, we suggest some practical solutions to these problems.

Solutions

Building better corpora

One of the most challenging issues to resolve is minimising the distance between those doing the data analysis and those researchers involved at other levels (e.g. field linguists). Part of the appeal of the nomothetic approach is the ease and cost-effectiveness in performing the analysis [14]. However, if the fundamental problems outlined in this paper are to be overcome, then there are a few solutions we can apply to this distance problem which involve improving the data quality. First, we want to increase the resolution of each individual variable. So, to take the previous example of consonant inventory size, the aim should be to report all accounts and not select one on the basis of prior theoretical assumptions. Having more data per variable will increase the statistical power for nomothetic studies. Second, minimising distance can be achieved by using multiple and, ideally, independent datasets that work together to build up mutually supporting evidence for or against a particular hypothesis. Different datasets can take the shape of those derived from different large-scale studies (e.g. Phoible [75] and WALS for phoneme inventory counts [55]), idiographic accounts of individual case studies and experimental data.

Thirdly, databases such as the WALS indicate linguistic norms for populations, but may not capture the variation within and between individuals. One solution is for the primary data to be raw text or recordings of real interactions between individuals [76] and for population-level features, such as grammatical rules, to be derived directly from these. While collecting adequate amounts of data of this kind is more difficult, and while it is not free of biases, it offers a richer source of information.

Furthermore, databases should be collected and coded with specific questions in mind, otherwise there is a risk that correlations

Table 3. Genes, collectivism, migration and ecology.

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	121.6578	25.6328	4.75	0.0002	*
5-HTTLPR short-short allele prevalence	-1.0900	0.2012	-5.42	0.0001	*
Gini coefficient	-0.4310	0.2694	-1.60	0.1292	
Contemporary pathogen prevalence	-0.6594	0.6995	-0.94	0.3599	
Historical pathogen prevalence	-5.5039	5.9007	-0.93	0.3648	
GDP	-0.0002	0.0003	-0.74	0.4717	
Net migration	4.1561	1.3598	3.06	0.0075	*
Biodiversity	0.1925	0.0718	2.68	0.0163	*
Minimum average temperature	2.7335	0.8841	3.09	0.0070	*
Mean growing season_calc	2.9013	1.3535	2.14	0.0478	
Minimum average temperature:					
Mean growing season_calc	-0.3352	0.1049	-3.20	0.0056	*

doi:10.1371/journal.pone.0070902.t003

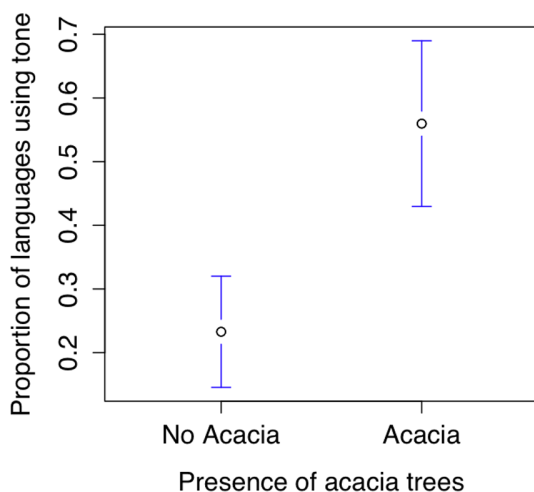


Figure 8. The proportion of tone languages within a country as a function of the presence of *Acacia nilotica*.
doi:10.1371/journal.pone.0070902.g008

could emerge due to biases in the original motivation for the database. For example, the database that was used to demonstrate a link between future tense and economic behaviour was designed to identify similarities between European languages, which also happen to be culturally related and relatively wealthy [36].

Model comparison

The correct null models to use when assessing cultural traits can be difficult to estimate, or unintuitive. As we shall demonstrate below, standard baselines of chance may not be conservative enough to eliminate spurious correlations. Rather than use random chance as a baseline, studies should compare competing hypotheses (as in [7]). Model comparison techniques allow researchers to test one model against another to see which better explains a particular distribution of data [77,78]. So, whereas standard regression techniques are able to tell you the amount of deviance explained by a particular model, they do not provide information about whether you should have a preference for one model over another given a particular set of data. Model comparison techniques are therefore useful summaries of the available information and are better viewed as inductive-style approaches that should be complementary to the hypothetico-deductive and falsificationist approaches more typically associated with the scientific process [72]. Model comparison can also be used to test linear versus non-linear assumptions.

Phylogenetic comparative methods

A simple, although conservative, test that controls for the relatedness of languages is to run the analysis within each language family (as in [1]). For example, the correlation between acacia trees and tonal languages is only significant for one language family, which is evidence against a causal relationship. However, more sophisticated methods are available. Studies of cultural traits have borrowed tools from biology to control for the non-independence of cultures [11]. Comparative methods include estimating the strength of a phylogenetic signal [49,79] and estimating the correlation between variables while controlling for the relatedness of observations [80–82]. For example, in the analyses above we found that speakers who take siestas have grammars with less verbal morphology. While experiments show that daytime naps affect procedural memory [83], which has been

Table 4. Tone and acacia trees by language family.

Family	Observations	t	p
Afro-Asiatic	29	-1.11	0.29
Austro-Asiatic	16	0.66	0.54
Austronesian	42	-1.73	0.17
Indo-European	30	0.27	0.81
Niger-Congo	64	4.99	0.000006
Nilo-Saharan	26	1.37	0.19
Sino-Tibetan	25	1.98	0.06
Trans-New Guinea	19	0.88	0.40

Results of t-tests for the relationship between linguistic tone and the presence of acacia trees within different language families. Columns indicate the language family, the number of languages used as observations in the test, the t-test statistic of the difference between tonal and non-tonal languages in terms of the presence of acacia trees and the probability value associated with that t-value.
doi:10.1371/journal.pone.0070902.t004

linked to morphological processing [84], the predictions run in the opposite direction to the results. However, doing the same analysis, but accounting for the relatedness of languages using a phylogenetic tree [80], this correlation disappears entirely ($r = 0.017$, $t = 0.13$, $p = 0.89$, see methods). This highlights the very different implications that can come out of nomothetic studies when considering the independence of the observations.

While phylogenetic methods are relatively new and phylogenetic reconstruction (see below) is computationally expensive, software for phylogenetic comparative methods is freely available (e.g. packages for R, [85–88]) and do not require intense computing power. The more limiting factor for studies of linguistic features is a lack of standard, high-resolution phylogenetic trees.

Other phylogenetic techniques have been used to reconstruct likely trees of descent from cultural data (e.g. [89–91]). These may also be useful as further steps for determining whether links between cultural traits discovered by nomothetic studies are robust. For example, apparent universals in the distribution of linguistic structural features may actually be underpinned by lineage-specific trends [92].

Causal graphs

Our analyses above suggests that cultural features are linked in complex ways, making it difficult to know what to control for in a specific study and potentially casting doubt on the value of nomothetic approaches. However, we see nomothetic studies as a useful tool for exploring complex adaptive systems. One change to the approach which could offer better resistance to the problems above would be to move away from trying to explain the variance in a single variable of interest towards analysing networks of interacting variables.

One method that could aid this type of analysis is the construction of causal graphs from large datasets [15]. While mediation analyses are often used to assess the causal relationship between a small number of variables [4], recent techniques are designed to handle high-dimensional data. We applied this technique to many of the variables in the study above. Figure 4 shows the most likely directed, acyclic graph that reflects the best fit to the relationships between the variables. We emphasise that this graph should be interpreted as a useful visualisation and as a hypothesis-generating exercise rather than representing proof of causation between variables.

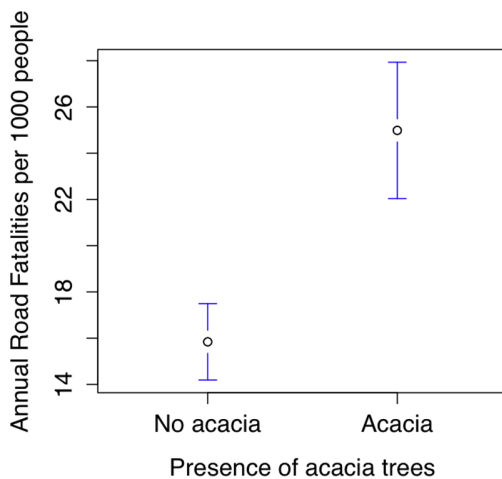


Figure 9. The mean number of annual road fatalities per 100,000 people within a country as a function of the presence of *Acacia nilotica*.

doi:10.1371/journal.pone.0070902.g009

Some interesting relationships emerge. First, some elements make intuitive sense, like the contemporary pathogen prevalence relying on the historical pathogen prevalence and the Gini coefficient (the balance between rich and poor within a country). Also, environmental variables like the number of frost days, mean growing season and mean temperature are linked.

More importantly, while the initial analysis above finds a direct correlation between linguistic diversity and road fatalities, even controlling for many factors, the causal graph analysis suggests that linguistic diversity and road fatalities are not causally linked. Instead, linguistic diversity is affected by demographic variables such as population size and density while road fatalities are affected by economic indicators such as GDP and the Gini coefficient. Similarly, the analysis suggests that tonal languages and the presence of acacia trees are not causally linked.

While the causal graph mainly provides evidence against some of the correlations above, it may also suggest interesting areas of further investigation. Interestingly, the causal graph suggests that collectivism is not directly linked with the genetic factors implicated by [4], but the relationship is mediated by (current) migration patterns. While speculative, it would be interesting to test the hypothesis that the distribution of genetic factors that are correlated with collectivism emerged by a process of selective migration (although see [93]). For example, the genotype that correlates with more collectivist countries is associated with a greater risk of depression under stress [29], so perhaps this gene came under selection in harsher climates. Indeed, we find some support for this idea, since adding environmental variables improves the fit of the model predicting the distribution of genotypes (compared to [4], see methods section). In this way, causal graph analyses may be a useful additional tool that can be used to explore relationships between complex adaptive variables such as cultural traits. Since the range of hypotheses suggested by inductive approaches can be very large, methods such as causal graphs can point to fruitful hypotheses to develop with more conventional approaches such as experiments.

Conclusion

Due to increasingly accessible data and analysis methods, there has been a recent rise in studies that use large-scale cross-cultural

Table 5. Traffic accidents and acacia trees.

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	18.5610	1.8427	10.07	<0.0001	*
Presence of acacia trees	4.9211	1.5106	3.26	0.0014	*
Length of road network	-0.0067	0.0038	-1.77	0.0790	
Greenberg diversity index	10.3095	2.1952	4.70	<0.0001	*
Nominal GDP	-0.0000	0.0000	-1.30	0.1949	
Distance from equator	-0.1524	0.0414	-3.68	0.0003	*
Population size	-0.0000	0.0000	-1.09	0.2781	
Population density	-0.0038	0.0010	-3.65	0.0004	*

A linear regression predicting the levels of road fatalities using presence of acacia trees, km of road, greenberg diversity index, nominal GDP, absolute latitude, population size and population density.

doi:10.1371/journal.pone.0070902.t005

databases to demonstrate correlations between cultural and demographic variables. While these studies may be useful for generating hypotheses and fostering interdisciplinary work, there are also problems which mean that they may have little explanatory power [14]. One of these problems is the relatedness of cultural groups and the correlated inheritance of cultural traits (Galton's problem). In this paper we illustrate the scale of the problem by demonstrating a chain of correlations between a diverse set of cultural traits. The probability of a spurious correlation between any two cultural traits is higher than is sometimes appreciated by researchers, the media and the general public.

We suggest four ways of addressing the problem of spurious correlations. First, better data will reduce the likelihood of correlations generated by noise. Secondly, we suggest that null models should be derived from alternative hypotheses rather than random chance. Thirdly, we encourage the development of phylogenetic techniques that account for the relatedness of cultures. Finally, we suggest moving from a paradigm of trying to explain the link between two variables towards explaining networks of interacting variables.

Although the explanatory power of these studies is weak, the appearance of rigour in the correlational analysis gives the related hypotheses credibility. Given the potential implications on policy for some cultural phenomena, conclusions from nomothetic studies could have negative effects. Researchers and reviewers should be cautious when evaluating approaches which link variables that are related by descent.

Materials and Methods

Here we describe the data and analyses used to demonstrate the spurious correlations between cultural variables discussed above.

Linguistic diversity and traffic accidents

The first analysis compared the linguistic diversity of a country to the number of fatal traffic accidents. The analysis contained data from 117 countries. A multiple regression was carried out

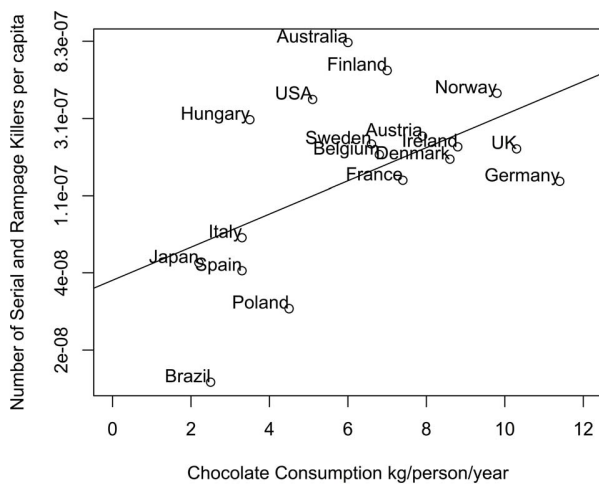


Figure 10. Chocolate consumption per capita (kg) as a function of the log number of serial and rampage killers per capita since 1900.

doi:10.1371/journal.pone.0070902.g010

with the Greenberg diversity index [94] as the dependent variable and the following independent variables: Road fatalities per 100,000 inhabitants per year [95]; population size [94]; population density [96]; nominal GDP, per-capita GDP [97]; net migration rate [98]; absolute latitude; absolute longitude; whether the country was inside or outside of Africa (a strong predictor of road fatalities) and total length of road network [99]. The fit of the model is improved by adding road fatalities after entering all other variables (RSS = 7.4781, $F(106,1) = 8.9$, $p = 0.0035$). Model adjusted $R^2 = 0.23$, $F(106,10) = 4.56$, $p < 0.0001$. Road fatalities coefficient = 0.012, $r = 0.45$, data available in Supporting Information S1, file S1_01.csv).

Siestas and morphological complexity

Countries with cultures of taking afternoon naps [100] are less morphologically complex, as measured by the mean number of grammatical categories a verb can take [101] ($n = 137$, $t = 3.47$, $p = 0.001$, see figure 5, Data available in Supporting Information S1, file S1_02.csv). Note that countries in Asia, Europe and South America take daytime naps. To test whether this is affected by Galton's problem, the language classifications for 127 languages were retrieved from the Ethnologue [102] and used to generate a phylogenetic tree (using H. Bibiko's AlgorithmTreeFromLabels program [103]). Without the phylogenetic tree, the correlation is significant (logit model: $r = -0.36096$, $z = -2.755$, $p = 0.00586$). To account for the phylogenetic tree, a generalised estimating equations test was run with binomial response distribution [80,85] (comparison suggested by chapter 7.7 of [88]). In this case, the correlation disappears ($r = 0.017$, $t = 0.13$, $p = 0.9$, $dfP = 63.2$, estimated scale parameter = 1.17, data available in Supporting Information S1, file S1_02b.zip).

Extramarital sex and population size

Extramarital sex is correlated with population size ($r = -0.54$, $df = 31$, $p = 0.001$), data from [104]. A regression using population size as the dependent variable and independent variables including population density and four measures of sexual attitudes showed that extramarital sex frequency was the best predictor after population density (see figure 6 and table 1, adjusted $R^2 = 0.38$, F

(26,5) = 4.84, $p = 0.003$, Data available in Supporting Information S1, file S1_03.csv).

Migration and Collectivism

Cultural values of collectivism are related to the prevalence of an allele of the serotonin transporter functional polymorphism 5-HTTLPR [4]. The original study used a linear regression with a measure of a country's collectivism as the dependent variable and independent variables including the prevalence of the 5-HTTLPR short-short allele, GDP, Gini coefficient and measures of pathogen prevalence. We replicated exactly the original finding that prevalence of the short-short allele is a significant predictor of collectivism (coefficient = -0.85 , $t = -2.94$, $p = 0.0079$). Adding contemporary migration levels [98] shows that migration levels are a significant predictor ($n = 28$, $r = 0.42$, see figure 7 and table 2) and improves the fit of the model (RSS difference = 1334, $F(22,1) = 10.3$, $p = 0.004$, adjusted $R^2 = 0.73$, data is available in Supporting Information S1, file S1_04.csv). Higher levels of collectivism (lower levels of individualism) correlate with lower migration rates.

Furthermore, we find that adding environmental variables (biodiversity [105], mean minimum annual temperature [106] and mean growing season [105] improves the fit of the model on top of the contribution from migration (RSS difference = 1318.8, $F(16,4) = 4.1745$, $p = 0.017$, adjusted $R^2 = 0.83$, see table 3).

Tone and Acacia Trees

Countries in which the acacia tree *Acacia nilotica* grows [107] were compared with countries which include tone languages (languages that use "pitch patterns to distinguish individual words or the grammatical forms of words", [108]). Acacia trees and tone languages (simple or complex) co-occur with a probability greater than chance (617 languages in 114 countries, χ^2 with Yates' continuity correction = 47.1, $df = 1$, $p < 0.0001$, see also [109], data available in Supporting Information S1, file S1_05.csv). The proportion of tonal (vs. non-tonal) languages in a country is significantly higher if that country has acacia trees (mean proportion of languages using tone in countries with acacia trees = 55.9% (41 countries), without acacia trees = 23.3% (73 countries), $t = 4.2$, $df = 76$, $p = 0.00007$, see figure 8). The proportion of languages with linguistic tone in a country predicts the presence of acacia trees, even when controlling for latitude (linear model, tone coefficient = 0.39, $t = 3.77$, $p = 0.0002$).

We can run an analysis of the relationship between tone and acacia trees within each language family. Enough data and variance was available for 8 language families (see table 4). The relationship was only significant for languages from the Niger-Congo family.

Acacia Trees and traffic accidents

Countries in which the acacia tree *Acacia nilotica* grows [107] have higher incidences of road fatality [95] (see figure 9, mean road fatalities per 100,000 inhabitants per year in countries without acacia trees = 15.84, mean in countries with acacia trees = 24.98654, $df = 85$, $p = 0.0000006$). A linear regression predicting the levels of road fatalities using presence of acacia trees, km of road [99], greenberg diversity index [94], nominal GDP [97], absolute latitude, population size [94] and population density [96] (see table 5), shows that the presence of acacia trees is a significant predictor (adjusted $R^2 = 43.1\%$, data available in Supporting Information S1, file S1_6.csv).

To test the geographic relatedness of countries, the distance between each country in the sample was calculated (great circle

distance from the center of each country) to produce a geographic distance matrix. Similar distance matrices were made for the GDI and road fatalities variables (absolute difference between countries). A Mantel test was used to calculate the probability of a correlation between GDI and road fatalities ($r=0.22$, $p=0.000001$, one million permutations). This remained significant when controlling for geographic distance with a partial Mantel test ($r=0.22$, $p=0.000001$, although see [48] for problems with Mantel tests).

Chocolate consumption and serial killers

We take five variables from Wikipedia: the number of Nobel prizes awarded by country of recipient (and the population of that country) [110]; The nominal gross domestic product (GDP) per capita [111]; the number of road fatalities per 10,000 population [112]; the number of serial killers since 1900 [113] and the number of rampage killers since 1900 [114]. The average annual temperature was obtained [105]. Data on the average IQ of the populations of different countries were obtained from [115]. The collected data is available in the supporting materials (Supporting Information S1, file S1_07.csv).

We replicated the finding from [9] that chocolate consumption per capita correlates with the number of Nobel laureates per capita ($r=0.73$, $p=0.00007$). However, a linear regression controlling for per-capita GDP and mean temperature found that chocolate consumption was not a significant predictor of the number of Nobel laureates ($F(1,19)=3.6$, $p=0.07$). Countries with higher GDP and lower mean temperatures correlate with higher Nobel laureates per capita ($r=0.7$, -0.6 , $p=0.0002,0.0016$). Furthermore, the average IQ of a country did not correlate with chocolate consumption ($r=0.27$, $p=0.21$). Additionally, for 18 countries where data was available, the level of chocolate consumption per capita is significantly correlated with the (log) number of serial killers and rampage killers per capita ($r=0.52$, $p=0.02$, see figure 10). We assume that there is no causal link here. Also, we found that the number of road fatalities per 100,000 inhabitants per year correlates with the number of Nobel Laureates ($r=-0.55$, $p=0.0066$), which we also assume has no causal link.

References

- Lupyan G, Dale R (2010) Language structure is partly determined by social structure. *PLoS ONE* 5: e8559.
- Chen MK (2013) The effect of language on economic behavior: Evidence from savings rates, health behaviors, and retirement assets. *American Economic Review* 103: 690–731.
- Atkinson QD (2011) Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa. *Science* 332: 346–349.
- Chiao JY, Blizinsky KD (2010) Culture-gene coevolution of individualism-collectivism and the serotonin transporter gene. *Proceedings of the Royal Society B: Biological Sciences* 277.
- Way BM, Lieberman MD (2010) Is there a genetic contribution to cultural differences? collectivism, individualism and genetic markers of social sensitivity. *Social Cognitive and Affective Neuroscience* 5: 203–211.
- Ember C, Ember M (2007) Climate, ecomiche, and sexuality: Influences on sonority in language. *American Anthropologist* 109: 180–185.
- Dediu D, Ladd D (2007) Linguistic tone is related to the population frequency of the adaptive haplogroups of two brain size genes, ASPM and Microcephalin. *Proceedings of the National Academy of Sciences* 104: 10944.
- Hay J, Bauer L (2007) Phoneme inventory size and population size. *Language* 2: 388–400.
- Messerli FH (2012) Chocolate consumption, cognitive function, and nobel laureates. *New England Journal of Medicine* 367: 1562–1564.
- Naroll R (1961) Two solutions to galton's problem. *Philosophy of Science* 28: 15–39.
- Levinson S, Gray R (2012) Tools from evolutionary biology shed new light on the diversification of languages. *Trends in Cognitive Sciences* 16: 167–173.
- Nettle D (2009) Ecological influences on human behavioural diversity: a review of recent findings. *Trends in ecology & evolution* 24: 618–624.

Causal graphs

The data from the studies above were aggregated over countries and combined into a single dataset. We used the PC algorithm [116] as implemented in the R package pcalg [117] to compute the most likely directed acyclic graph of relationships between variables. The algorithm has a parameter that determines the threshold at which links should be included. The results come from using the smallest threshold that included all the variables in a single connected component. We note that the exact causal links that are selected are sensitive to this parameter and to different subsets of the data. Therefore, we suggest that this method is only an exploratory tool rather than a formal proof of relationships. We look forwards with anticipation to the development of this tool.

Supporting Information

Supporting Information S1 Contains: S1_01.csv: Data on road fatalities, linguistic diversity and demographic variables for countries. **S1_02.csv:** Data on Siestas and morphological complexity. **S1_02b/AlgTree_AJJP_Ethno.nwk:** Phylogenetic tree of languages. **S1_02b/s.csv:** Data on Siestas and morphological complexity. **S1_02b/Siesta_PhyloLogit.r** R script for running phylogenetic generalised estimating equations test. **S1_03.csv:** Data on population size and extramarital sex frequency. **S1_04.csv:** Data on genetic correlates of collectivism and migration. **S1_05.csv:** Data on Tone languages and Acacia trees. **S1_6.csv:** Data on Acacia tree and traffic accidents. **S1_07.csv:** Data on Chocolate consumption and serial killers. (ZIP)

Acknowledgments

We thank Gary Lupyan and one anonymous reviewer for excellent comments. Thanks also to Michael Dunn, Sarah Graham, Elizabeth Irvine, Eric Johnstone, Andrew Oh-Willeke and the readers of *A Replicated Typo* for comments and discussion.

Author Contributions

Conceived and designed the experiments: SR JW. Performed the experiments: SR JW. Analyzed the data: SR JW. Contributed reagents/materials/analysis tools: SR JW. Wrote the paper: SR JW.

23. Wichmann S, Rama T, Holman E (2011) Phonological diversity, word length, and population sizes across languages: The ASJP evidence. *Linguistic Typology* 15: 177–197.
24. Moran S, McCloy D, Wright R (2012) Revisiting population size vs. phoneme inventory size. *Language* 88: 877–893.
25. Donohue M, Nichols J (2011) Does phoneme inventory size correlate with population size. *Linguistic Typology* 15: 161–170.
26. Trudgill P (2004) Linguistic and social typology: The austronesian migrations and phoneme inventories. *Linguistic Typology* 8: 305–320.
27. Jaeger TF, Graff P, Croft W, Pontillo D (2011) Mixed effect models for genetic and areal dependencies in linguistic typology. *Linguist Typol* 15: 281–319.
28. Fought JG, Munroe RL, Fought CR, Good EM (2004) Sonority and climate in a world sample of languages: Findings and prospects. *Cross-cultural research* 38: 27–51.
29. Caspi A, Sugden K, Moffitt TE, Taylor A, Craig IW, et al. (2003) Influence of Life Stress on Depression: Moderation by a Polymorphism in the 5-HTT Gene. *Science* 301: 386–389.
30. Desideri G, Kwik-Urbe C, Grassi D, Necozone S, Ghiadoni L, et al. (2012) Benefits in cognitive function, blood pressure, and insulin resistance through cocoa flavanol consumption in elderly subjects with mild cognitive impairment: novelty and significance of the cocoa, cognition, and aging (cocoa) study. *Hypertension* 60: 794–801.
31. Pritchard C (2012) Does chocolate make you clever? *BBC News Magazine Online* 19/11/2012 <http://www.bbc.co.uk/news/magazine-20356613>. Accessed 18/04/2013.
32. Husten L (2012) Chocolate and nobel prizes linked in study. *Forbes: Pharma & Healthcare Online* 10/10/2012 <http://www.forbes.com/sites/larryhusten/2012/10/10/chocolate-and-nobel-prizes-linked-in-study/> Accessed 18/04/2013.
33. Joelsing F (2012) Eat chocolate, win the nobel prize? *Reuters US Online* 10/10/2012 <http://www.reuters.com/article/2012/10/10/us-chocolate-nobels-idUSBRE8991SS20121010> Accessed 18/04/2013.
34. Liberman P (2012) Cultural diffusion and the whorfian hypothesis. *Language Log*, Posted February 12, 2012, accessed May 21, 2012 <http://languageblogdcpennedu/nll/?p=3764>.
35. Pullum GK (2012) Keith chen, whorfian economist. *Language Log*, Posted February 9, 2012, accessed May 21, 2012 (<http://languageblogdcpennedu/nll/?p=3756>).
36. Dahl O (2013) Stuck in the futureless zone. *Diversity Linguistics comment* Posted 03/09/2013 Accessed 18/04/2013 <http://dlchypothesesorg/360>.
37. Berreby D (2012) Obese? smoker? no retirement savings? perhaps it's because of the language you speak. *Big Think* :February 5, 2012. <http://bigthink.com/ideas/42306>.
38. Fellman B (2012) Speaking and saving. *Yale Alumni Magazine* January 1.
39. Chen K (2012) Could your language affect your ability to save money? *TEDGlobal* 2012 Online http://www.ted.com/talks/keith_chen_could_your_language_affect_your_ability_to_save_money.html Accessed 18/04/2013.
40. Keating JE (2012) Tomorrow, we save. *Foreign Policy* 01/10/2012 Online http://www.foreignpolicy.com/articles/2012/08/13/tomorrow_we_save Accessed 18/04/2013.
41. Bowler T (2013) Why speaking English can make you poor when you retire. *BBC News: Business* Posted 23/02/2013, accessed 19/04/2013. Online <http://www.bbc.co.uk/news/business-21518574>.
42. Simonton DK (1975) Galton's problem, autocorrelation, and diffusion coefficients. *Cross-Cultural Research* 10: 239–248.
43. Beckner C, Blythe R, Bybee J, Christiansen MH, Croft W, et al. (2009) Language is a complex adaptive system: Position paper. *Language Learning* 59: 1–26.
44. Ross MH, Homer E (1976) Galton's problem in cross-national research. *World Politics* 29: 1–28.
45. Gray RD, Jordan FM (2000) Language trees support the express-train sequence of austronesian expansion. *Nature* 405: 1052–1055.
46. Terrell JE, Hunt TL, Gosden C (1997) Human diversity and the myth of the primitive isolate. *Current Anthropology* 38: 155–195.
47. Collard M, Shennan SJ, Tehrani JJ (2006) Branching, blending, and the evolution of cultural similarities and differences among human populations. *Evolution and Human Behavior* 27: 169–184.
48. Nunn CL, Mulder MB, Langley S (2006) Comparative methods for studying cultural trait evolution: A simulation study. *Cross-Cultural Research* 40: 177–209.
49. Pagel M (1999) Inferring the historical patterns of biological evolution. *Nature* 401: 877–884.
50. Graur D, Zheng Y, Price N, Azevedo RB, Zufall RA, et al. (2013) On the immortality of television sets: function in the human genome according to the evolution-free gospel of encode. *Genome biology and evolution*.
51. Bennett CM, Baird AA, Miller MB, Wolford GL (2011) Neural correlates of interspecies perspective taking in the post-mortem atlantic salmon: an argument for proper multiple comparisons correction. *Journal of Serendipitous and Unexpected Results* 1: 1–5.
52. Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, et al. (2013) Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* 14: 365–376.
53. Todd MT, Nystrom LE, Cohen JD (2013) Confounds in multivariate pattern analysis: Theory and rule representation case study. *NeuroImage* 77: 157–165.
54. Sproat R (2010) Ancient symbols, computational linguistics, and the reviewing practices of the general science journals. *Computational Linguistics* 36: 585–594.
55. Maddieson I (2011) Consonant inventories. In: Dryer MS, Haspelmath M, editors, *The World Atlas of Language Structures Online*. Accessed 18/04/2013, Munich: Max Planck Digital Library. URL <http://wals.info/feature/1A>.
56. Avram MLZ (2008) A phonological description of wichí: the dialect of misión la paz, salta, argentina. *Masters Theses and Doctoral Dissertations* : 152.
57. Tovar A (1981) *Relatos y diálogos de los matacos: Seguidos de una gramática de su lengua*. Madrid: Ediciones Cultura Hispánica del Instituto de Cooperación Iberoamericana.
58. Claesson K (1994) A phonological outline of mataco-noctenes. *International Journal of American Linguistics* 60: 1–38.
59. Taleb NN (2012) *Antifragile: things that gain from disorder*. Random House Incorporated.
60. Cohen J (1988) *Statistical Power Analysis for the Behavioral Sciences*, second edition. Lawrence Erlbaum, NJ.
61. Fanelli D (2010) positive results increase down the hierarchy of the sciences. *PLoS one* 5: e10068.
62. CMS collaboration and others (2012) Observation of a new boson with a mass near 125 gev. *CMS physics analysis summary CMSPAS-HIG-12 20*.
63. ATLAS collaboration and others (2012) Observation of an excess of events in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *ATLAS-CONF-2012-093*.
64. Yanny B, Rockosi C, Newberg HJ, Knapp GR, Adelman-McCarthy JK, et al. (2009) *Segue: A spectroscopic survey of 240,000 stars with g = 14–20*. *The Astronomical Journal* 137: 4377.
65. Marcus G (2013) Steamrolled by big data. *The New Yorker Elements Blog* Posted 03/04/2013, accessed 19/04/2013. Online: <http://www.newyorker.com/online/blogs/elements/2013/04/steamrolled-by-big-data.html>.
66. Rogoff K, Reinhart C (2010) Growth in a time of debt. *American Economic Review* 100: 573–78.
67. Bivens J, Irons J (2010) Government debt and economic growth. *Economic policy institute Report: Budget Taxes and Public Investment* 26/07/2010. Online <http://www.epi.org/publication/bp271/> Accessed 18/04/2013.
68. Herndon T, Ash M, Pollin R (2013) Does high public debt consistently stie economic growth? a critique of reinhart and rogoff. *Political Economy Research Institute Workingpaper series* 22.
69. United States Senate Committee on the Budget (2011) The case for growth: Sessions lists benefits of discretionary cuts. Online 15/03/2011 <http://www.budgetsenategov/republican/public/indexcfm/2011/3/the-case-for-growth-sessions-lists> Accessed 18/04/2013.
70. Yglesias M (2013) Is the reinhart-rogoff result based on a simple spreadsheet error? *Slate magazine Moneybox blog* Online, 16/04/2013 http://www.slate.com/blogs/moneybox/2013/04/16/reinhart_rogoff_coding_error_austerity_policies Accessed 18/04/2013.
71. Dedić D (2013) Genes: Interactions with language on three levels: inter-individual variation, historical correlations and genetic biasing. In: *The Language Phenomenon*, Springer: 139–161.
72. Gelman A, Shalizi CR (2012) *Philosophy and the practice of bayesian statistics*. *British Journal of Mathematical and Statistical Psychology*.
73. Nettle D (1998) Explaining global patterns of language diversity. *Journal of anthropological archaeology* 17: 354–374.
74. Murray EJ (1965) *Sleep, dreams, and arousal*. New York: Appleton-Century-Crofts.
75. Moran S (2012) *Phonetics Information Base and Lexicon*. Ph.D. thesis, University of Washington.
76. Levinson S (2006) On the human interaction engine. In: Enfield N, Levinson S, editors, *Roots of Human Sociality: Culture, Cognition and Human Interaction*, Oxford: Berg: 39–69.
77. Congdon P (2005) *Bayesian models for categorical data*. Wiley.
78. Alston C, Kuhnert P, Choy SL, McVinish R, Mengersen K (2005) Bayesian model comparison: Review and discussion. *International Statistical Institute*, 55th session.
79. Fritz SA, Purvis A (2010) Selectivity in mammalian extinction risk and threat types: a new measure of phylogenetic signal strength in binary traits. *Conservation Biology* 24: 1042–1051.
80. Paradis E, Claude J (2002) Analysis of comparative data using generalized estimating equations. *Journal of Theoretical Biology* 218: 175–185.
81. Martins EP, Hansen TF (1997) Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *American Naturalist*: 646–667.
82. Pagel MD (1992) A method for the analysis of comparative data. *Journal of theoretical Biology* 156: 431–442.
83. Backhaus J, Junghanns K (2006) Daytime naps improve procedural motor memory. *Sleep Medicine* 7: 508–512.
84. Ullman M (2005) A cognitive neuroscience perspective on second language acquisition: The declarative/procedural model. In: Sanz C, editor, *Mind and Context in Adult Second Language Acquisition: Methods, Theory, and Practice*, Georgetown University Press: 141–178.
85. Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20: 289–290.
86. Orme D, Freckleton R, Thomas G, Petzoldt T, Fritz S, et al. (2012) *caper: Comparative Analyses of Phylogenetics and Evolution in R*. URL <http://>

- CRAN.R-project.org/package=caper.R package version 0.5. Accessed 18/04/2013.
87. Harmon L, Weir J, Brock C, Glor R, Challenger W, et al. (2009) geiger: Analysis of evolutionary diversification. URL <http://CRAN.R-project.org/package=geiger>. R package version 1.3–1. Accessed 18/04/2013.
 88. Nunn C, editor (2013) The AnthroTree website. Online <http://numn.rc.fas.harvard.edu/groups/pica/>. Accessed 18/04/2013.
 89. Bouckaert R, Lemey P, Dunn M, Greenhill SJ, Alekseyenko AV, et al. (2012) Mapping the origins and expansion of the indo-european language family. *Science* 337: 957–960.
 90. Gray RD, Drummond AJ, Greenhill SJ (2009) Language phylogenies reveal expansion pulses and pauses in pacific settlement. *Science* 323: 479–483.
 91. Mace R, Jordan FM (2011) Macro-evolutionary studies of cultural diversity: A review of empirical studies of cultural transmission and cultural adaptation. *Philosophical Transactions of the Royal Society B: Biological Sciences* 366: 402–411.
 92. Dunn M, Greenhill SJ, Levinson SC, Gray RD (2011) Evolved structure of language shows lineage-specific trends in word-order universals. *Nature* 473: 79–82.
 93. Eisenberg DT, Hayes MG (2011) Testing the null hypothesis: comments on culture-gene coevolution of individualism–collectivism and the serotonin transporter gene. *Proceedings of the Royal Society B: Biological Sciences* 278: 329–332.
 94. Lewis MP, editor (2009) *Statistical Summaries*, Dallas, Tex.: SIL International, volume *Ethnologue: Languages of the World*, Sixteenth edition.
 95. World Health Organization Injuries and Violence Prevention Dept (2002) *The injury chart book: A graphical overview of the global burden of injuries*. World Health Organization.
 96. Wikipedia. List of sovereign states and dependent territories by population density. accessed 18/04/2013. URL http://en.wikipedia.org/wiki/List_of_sovereign_states_and_dependent_territories_by_population_density.
 97. The World Bank. World development indicators. accessed 18/04/2013. URL <http://data.worldbank.org/indicator/NY.GDP.MKTP.CD/countries>.
 98. Central Intelligence Agency. The world factbook: Net migration rate. accessed 18/04/2013. URL <https://www.cia.gov/library/publications/the-world-factbook/rankorder/2112rank.html>.
 99. Central Intelligence Agency (2012). The world factbook: Roadways. accessed 18/04/2013. URL <https://www.cia.gov/library/publications/the-world-factbook/rankorder/2085rank.html>.
 100. Wikipedia (2011). Siesta. accessed 01/05/2011. URL <http://en.wikipedia.org/wiki/Siesta>.
 101. Bickel B, Nichols J (2011) Inectional synthesis of the verb. In: Dryer MS, Haspelmath M, editors, *The World Atlas of Language Structures Online*, Munich: Max Planck Digital Library. Accessed 18/04/2013. URL <http://wals.info/chapter/22>.
 102. Gordon R (2005) *Ethnologue: Languages of the World*, 15th Edition. SIL International.
 103. Bibiko HJ (2009) Program for creating nexus files from classificatory language labels (version 1.0) Online <http://email.eva.mpg.de/~wichmann/software.htm>. Accessed 18/04/2013.
 104. Murdock G, White D (1969) Standard cross-cultural sample. *Ethnology* 8: 329–369.
 105. Mitchell T, Hulme M, New M (2002) Climate data for political areas. *Area* 34: 109–112.
 106. Caldecott J, Jenkins M, Johnson T, Groombridge B (1994) Priorities for conserving global species richness and endemism. In: Collins NM, editor, *World Conservation Monitoring Centre, Biodiversity Series No. 3*, World Conservation Press, Cambridge, UK. p. 17.
 107. Crop Protection Compendium (2008). *Acacia confusa*. online, accessed 18-04-2011. <http://www.cabi.org>. URL <http://www.cabi.org>.
 108. Maddieson I (2011) Tone. In: Dryer MS, Haspelmath M, editors, *The World Atlas of Language Structures Online*, Munich: Max Planck Digital Library. URL <http://wals.info/chapter/13>.
 109. Geraint S (2011) Language evolution and the acacia tree. *SpecGram CLXII*.
 110. Wikipedia. List of countries by nobel laureates per capita. accessed 22/04/2013. URL http://en.wikipedia.org/wiki/List_of_countries_by_Nobel_laureates_per_capita.
 111. Wikipedia. List of countries by gdp (nominal) per capita. accessed 22/04/2013. URL [http://en.wikipedia.org/wiki/List_of_countries_by_GDP_\(nominal\)_per_capita](http://en.wikipedia.org/wiki/List_of_countries_by_GDP_(nominal)_per_capita).
 112. Wikipedia. List of countries by traffic-related death rate. accessed 22/04/2013. URL http://en.wikipedia.org/wiki/List_of_countries_by_traffic-related_death_rate.
 113. Wikipedia. List of serial killers by number of victims. accessed 22/04/2013. URL http://en.wikipedia.org/wiki/List_of_serial_killers_by_number_of_victims.
 114. Wikipedia. List of rampage killers. accessed 22/04/2013. URL http://en.wikipedia.org/wiki/List_of_rampage_killers.
 115. Lynn R, Vanhanen T (2002) *IQ and the wealth of nations*. Praeger Publishers.
 116. Spirtes P, Glymour CN, Scheines R (2000) *Causation, Prediction, and Search*. MIT press. 2nd Edition.
 117. Kalisch M, Mächler M, Colombo D, Maathuis MH, Bühlmann P (2012) Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software* 47: 1–26.

5.4 Conclusion

There are several challenges with using large-scale, cross-cultural analyses to investigate the relationship between context and language structure. Some of these challenges include: (i) finding a candidate dataset that allows us to operationalise context, (ii) controlling for confounding factors, and (iii) systematically testing hypotheses about the relationship between context and language structure.

First of all, finding a candidate dataset is beset by the general issue of an *inverse sample size problem*: the signal-to-noise ratio is likely to be large in cross-cultural datasets, as these tend to be incomplete, complex, and based on inconsistent criteria. Secondly, this chapter showed that any attempt to demonstrate a correlation needs to overcome the twin problems of shared descent between cultural groups (Galton's Problem) and the diffusion of cultural traits due to geographical proximity. Lastly, even if we overcome these two major problems, the focus of the study becomes about predicting typological distributions, and not about establishing whether there is a causal relationship.

Experiments on the other hand are well placed for this type of investigation: they allow us to probe whether or not a variable, when removed from its natural environment, has a causal effect on another variable (Nunan, 1992). As demonstrated in this thesis, experimental methods can directly test the hypothesis that context is a constraint on the structure of language, through framing what is and is not informative for discrimination. Relating experimental results to typological patterns is a necessary next step, and it is only when this is done can we adequately address the question of why the effect of context is or is not manifest in typological distributions.

Chapter 6

Conclusion

Why does the structure and organisation of language appear to be context-dependent? One answer is that language did not evolve for the purposes of communication (Chomsky, 2002; Berwick & Chomsky, 2016). Under this account, properties such as ambiguity are viewed as an impediment to communication, with contextual information being utilised to solve these problems. Another perspective flips the argument on its head and claims the exact opposite: language is context-dependent because it is communicatively efficient (Piantadosi et al, 2012). Irrespective of one's stance on these matters, what both of these proposals highlight is that understanding the relationship between context, cognition, and communication is fundamental in any account of how language evolved.

The first step is to arrive at a tractable definition of context. The argument put forward by this thesis is that context is best thought of as a frame of interpretation. Context in this sense consists of a figure (the target of interpretation), a ground (the immediate information brought to the act of interpretation), and a background (prior knowledge derived from previous frames). When interpreting a linguistic utterance, such as *James passed the port*, contextual information fills in the expressive gaps by highlighting the relevant information (such as our knowledge that *port* is a type of fortified wine and James is at a dinner party) and backgrounding irrelevant information (such as our knowledge that *port* is also a type of harbour). This definition of context as a frame of interpretation is important because it allows us to investigate the relationship between language structure and context as a *discrimination pressure*: determining what is and is not informative in reducing uncertainty in interpretation.

Importantly, this link between language structure and context emerges via cultural evolutionary processes, where the discrimination pressure interacts with two other pressures inherent to cognition and communication: a *generalisation*

pressure, which arises across repeated exposures and refers to the discrepancy between what is currently predicted and incoming sensory data, and a *coordination pressure*, where speakers and hearers must align on shared frames of interpretation. Language adapts to these pressures by establishing a set of conventional form-meaning mappings that are both expressive and compressible. A key question this thesis set out to answer is: How can we investigate these pressures in a systematic manner?

The majority of this thesis was therefore devoted to a series of experiments which investigated the effects of context on the structure of language. Experiment 1 set the groundwork for the rest of the thesis and demonstrated three points. First, manipulations to the referential context can be considered a useful proxy for investigating the general effects of context as a frame of interpretation. Second, on a methodological front, Experiment 1 brought together insights from work in artificial language learning, reference games, iterated learning, and communication games to create a framework for systematically investigating the effects of context on language structure. Lastly, the experiment showed how simple manipulations to the referential context result in radically different communication systems, with languages gradually adapting to how context interacts with these pressures of discrimination, generalisation, and coordination.

Experiment 2 began by reversing the usual premise found in these experiments: rather than investigating how structure emerges in language, this experiment focused on the relationship between context and the loss of language structure. One of the motivating factors for investigating the loss of language structure comes from observations in the historical record: here, due to the effects of reinterpretation and chunking, individual constructions undergo a loss of compositional structure. Context is viewed as a key factor in instantiating this change by introducing ambiguity into how a form maps onto a meaning. To tackle this specific issue, Experiment 2 addressed the question of whether changes to the discrimination pressure interacted with a pressure to generalise: when context backgrounded a dimension (e.g., colour) an initially compositional language is predicted to lose this distinction when generalising to new, unseen referents. The findings somewhat supported this general prediction, but only when context interacted with a shape bias. Experiment 3 built upon these findings by controlling for a possible confound (i.e., counterbalancing for the number of features), using brighter coloured referents, and employing a larger sample. With these changes the shape bias disappeared, and was replaced by a general preference to encode colour (albeit at a lower overall effect size than in Experiment 2).

Experiments 2 and 3 tell us there is much to learn about the relationship

between the loss of structure and context. One tentative conclusion from these two experiments is that, in this specific instance at least, the effect of context is not very robust – and is mediated by interactions with other biases acting on maintaining or decreasing language structure. For instance, in Experiment 2 the saliency of shape was increased, and this led to languages being underspecified on colour, and in Experiment 3 the saliency of colour was increased, and this led to languages being underspecified on shape. Whatever the reason is for the increase in saliency, be it due to an unbalanced meaning space or some intrinsic property of the stimuli, underspecification only happens when it is functionally adequate for discriminating between referents in a context. The logical next step is to probe deeper into clarifying the size and robustness of these effects.

The idea that the effects of context are mediated by other biases and constraints permeates the thinking behind Experiment 4: we investigated how context-type and access to this contextual information interact in shaping the degree of signal autonomy. Conditions where manipulations to contextual predictability are reinforcing result in participants converging on similar solutions to the coordination pressure. In Shape-Different + Shared Context this meant the systems were consistently context-dependent and in Mixed + Unshared Context this resulted in more autonomous systems. When these manipulations to contextual predictability are conflicting, as was the case in Shape-Different + Unshared and Mixed + Shared, participants produce heterogeneous systems in terms of autonomy (some were autonomous, some were context-dependent). Predictable contexts allow speakers and hearers to coordinate on a shared system which relies less on linguistic information to convey the intended meaning; context and hearer inference fill in any expressive gaps. As contextual predictability decreases, speakers become increasingly reliant on the linguistic system for coordinating with hearers, using strategies that promote more autonomous signals.

There are plenty of opportunities for future work in this area. One possible extension is to manipulate the size of the referential context. The experiments in this thesis ranged from contexts consisting solely of a target and a distractor (Experiments 1, 2 and 3) to contexts with a target and three distractors (Experiment 4). In terms of identifying the intended meaning, a larger referential context has higher uncertainty than a smaller context, with a hearer needing to sift through more distractors. What is interesting about this example is that the reverse is true for the speaker: a larger referential context is more informative than a smaller context. For the largest possible context, where the number of objects is equal to the total number of referents, speakers have access to more information about the necessary distinctions which are globally required by the

linguistic system. By contrast, having to discriminate between a single target and distractor only reveals what is locally relevant for discrimination, and is therefore less informative for the speaker in discovering the optimal system for conveying the intended meaning. This tension between what is informative for the speaker versus what is predictable for the hearer is a promising avenue for future research.

The experimental pragmatics literature offers a few interesting extensions for investigating the effects of context on language structure. As an example, Frank & Goodman (2012) provide a series of manipulations to the context-type, systematically manipulating whether one, two, or all three of the distractors in a referential context share a feature with the target. More fine-grain manipulations of the referential context would allow us to stress test some of the claims made in this thesis. To illustrate, consider a set of distractors which share a particular feature in common (e.g., *red oval*, *red rectangle*, *red star*) versus a set of distractors which are maximally distinct from one another (e.g., *grey oval*, *red rectangle*, *yellow star*). Now, imagine the target is a *blue blob*: in both context-types the optimal signal is one which underspecifies and conveys either blue or blob. Yet we do not know whether speakers will underspecify in such situations (as opposed to using, say, a compositional signal) and neither do we know whether one context-type will favour the encoding of one dimension over another (e.g., encoding shape instead of colour). Furthermore, were speakers to underspecify more frequently in the maximally distinct context, then this would run counter to the predictions outlined in Chapter 3 (i.e., participants are more likely to underspecify when one of feature dimensions is backgrounded to a certain extent). Such manipulations are important for linking up predictions about short-term pragmatic reasoning with the long-term emergence of communication systems.

Manipulations to the contextual information shared between interlocutors (i.e., common ground) is also a prominent feature of experimental pragmatics (for reviews, see: Brennan, Galati & Kuhlen, 2010; Konopka & Brown-Schmidt, 2014). Yet, in contrast to Experiment 4, where our focus was on whether or not the speaker shared access to the context given to the hearer, common ground experiments have also looked at the opposite situation: where access to the context is modulated for the hearer (e.g., Horton & Keysar, 1996). Introducing additional manipulations of the shared context, such as providing speakers and hearers with access to different referents in an array, will offer further insights into the relationship between signal autonomy and contextual predictability.

Another possible extension is to use the framework presented in this thesis to help address outstanding questions in referential games. For instance, unlike material and scalar adjectives, which tend to be dependent on context, colour ad-

jectives are often used even when they are uninformative for discrimination (e.g., Sedivy, 2005; Arts et al., 2011; Rubio-Fernandez, 2016). Much of the current research into *Redundant Colour Adjectives* (RCA) has only looked at their use in natural languages. Artificial languages are well-placed to investigate under what conditions RCAs do and do not emerge in communication systems. Furthermore, investigating RCAs using artificial languages has the added advantage of resolving some of the issues raised in Chapter 3, especially in understanding the role colour saliency plays in these experiments.

These experiments also provide some insights into how the generalisation pressure relates to the emergence of language structure. In the case of Experiment 1, participants at the first generation were faced with a strong generalisation pressure, as it was extremely difficult for them to memorise all of the signal-referent mappings in the initial language. For Experiments 2, 3 and 4, the generalisation pressure was more explicit: we restricted the number of referents participants saw and investigated how unseen referents were expressed during communication. Future work should systematically delineate between these different types of generalisation pressures in communication games (but see: Cornish, 2010¹). This will allow us to see under what conditions these two types of generalisation pressure produce similar and different outcomes with respect to the emergence of language structure.

The experimental models in this thesis were idealised versions of communication, leaving out integral properties inherent to the day-to-day use of language. Much of the work presented comes from a tradition of starting from the perspective of learning (e.g., early iterated learning models: Kirby & Hurford, 2002; Kirby, Cornish & Smith, 2008) and gradually incorporating more aspects of communication into our models of language (Kirby et al., 2015; Tamariz & Kirby, 2016). Still, much work remains to be done in bridging the gap between useful abstraction and real-world constraints, especially in regards to the importance of turn-taking (Stivers et al., 2009; Levinson, 2016) and conversational repair mechanisms (Hayashi, Raymond & Sidnell, 2013; Dingemans, Torreira & Enfield, 2013). The present set of experiments were relatively simple in this respect. There was no real-time repair and turn-taking was either restricted to simple alternations (Experiments 1, 2 and 3) or non-existent (Experiment 4). Incorporating these aspects might diminish the impact of the referential context – as

¹Cornish (2010) contrasts two types of generalisation pressure, a memory bottleneck (memory limitations of a participant to learn a set of form-meaning mappings) and a data bottleneck (the number of form-meaning mappings a participant is exposed to in their training set), but this has only been investigated in iterated learning (diffusion chain) experiments and not the communication game framework employed in this thesis.

participants have recourse to a new source for reducing uncertainty about the intended meaning. Although it remains to be seen whether or not participants will privilege repair and turn-taking over information provided by the referential context.

The last major point is a methodological one. Chapter 5 highlighted the challenges of a cross-linguistic approach as a means of investigating the relationship between language structure, context, and other pressures found in cognition and communication. Due to the nature of culturally transmitted behaviours, which are characterised by inheritance and borrowing, cross-cultural datasets are highly susceptible to spurious correlations. When combined with difficulty of operationalising context, especially in regards to investigating causal effects on language structure, the chapter concluded by arguing for an experimental approach. But this is not to say these approaches should never be employed. There is plenty of scope for cross-linguistic correlations to kick-start the traditional scientific process and stress-test previous empirical findings (for further discussion on both these points, see Roberts, Winters & Chen, 2015). Furthermore, if we are to relate the experimental results of this thesis to the typological patterns observed in the world's 6-7000 languages, then this gap between methodological approaches needs to be bridged (see Lewis & Frank, 2016 for a useful demonstration as to how one might link experimental findings to typological patterns with regards to conceptual complexity). Ideally, statistical approaches can be combined with simulations in generating testable hypotheses, helping whittle down the parameter space to a level that is tractable for experimental research (for tentative first steps in this direction, see: Winter & Ardell, 2016).

In summary, context is considered a crucial component in both learning and using a language, yet its role in cultural evolutionary accounts of language structure was traditionally taken as a given (see Scott-Phillips, 2015 for a fuller discussion on the role of pragmatics in language evolution research). This thesis demonstrates that context plays an important part in the cultural evolution of language. The first step was to establish context as a frame of interpretation: this allowed us to describe the effects of context as stemming from a discrimination pressure that interacts with the pressures of generalisation and coordination. Second, the thesis highlighted the strengths of using an experimental approach – contrasting this with other means of investigation such as cross-linguistic correlations – and demonstrated its utility for systematically investigating the effects of context on language structure. Lastly, the results of these experiments show how context constrains language structure in predictable ways, opening up new avenues for exploring the cultural evolution of language.

Bibliography

Alston, C., Kuhnert, P., Choy, S. L., McVinish, R., & Mengersen, K. (2005). Bayesian model comparison: Review and discussion. *International Statistical Institute*, 55th session.

Altmann, E. G., Pierrehumbert, J. B., & Motter, A. E. (2011). Niche as a determinant of word fate in online groups. *PLoS ONE*, 6(5): e19009. doi: 10.1371/journal.pone.0019009.

Andersen, H. (1973). Abductive and deductive change. *Language*, 49: 765-793.

Arts, A., Maes, A., Noordman, L., & Jansen, C. (2011). Overspecification facilitates object identification. *Journal of Pragmatics*, 43: 361-374.

Atkinson, M., Kirby, S., & Smith, K. (2015). Speaker Input Variability Does Not Explain Why Larger Populations Have Simpler Languages. *PLoS ONE*, 10(6): e0129463. doi: 10.1371/journal.pone.0129463.

Atkinson, Q. D. (2011). Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa. *Science*, 332: 346-349. doi: 10.1126/science.1199295.

ATLAS collaboration and others. (2012). Observation of an excess of events in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *ATLAS-CONF-2012-093*.

- Attardo, S. (2016). Context as Relevance-Driven Abduction and Charitable Satisficing. *Frontiers in Psychology*, 7. doi: 10.3389/fpsyg.2016.00305.
- Avram, M. L. Z. (2008). *A phonological description of Wichí: the dialect of Misión la Paz, Salta, Argentina*. Masters Theses and Doctoral Dissertations: 152.
- Ay, N., Flack, J. C. & Krakauer, D. C. (2007). Robustness and complexity co-constructed in multimodal signaling networks. *Philosophical Transactions of the Royal Society B*, 362: 441-447.
- Backhaus, J., & Junghanns, K. (2006). Daytime naps improve procedural motor memory. *Sleep Medicine*, 7: 508-512. doi: 10.1016/j.sleep.2006.04.002.
- Baars, B. J. (1988). *A cognitive theory of consciousness*. Cambridge: Cambridge University Press.
- Baayen, R. H. (2008). *Analysing linguistic data: a practical introduction to statistics using R*. Cambridge: Cambridge University Press.
- Bach, K. (2012). Context dependence (such as it is). In M. Garcia-Carpintero & M. Kolbel (Eds.), *The Continuum Companion to the Philosophy of Language*: Chapter 7. London: Bloomsbury Publishing.
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure in mixed-effects models: Keep it maximal. *Journal of Memory and Language*, 68(3): 255-278.
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1): 1-48. doi: 10.18637/jss.v067.i01.

Bazire, M. & Brezillon, P. (2005). Understanding context before using it. In B. Kokinov A. Dey (Eds.), *Modeling and Using Context*: 29-40. Springer.

Beckner, C., et al. (2009). Language is a complex adaptive system. *Language Learning*, 59(s1): 1-26.

Bell, A. (1984). Language style as audience design. *Language in Society*, 13: 145-204. doi: 10.1017/S004740450001037X.

Bennett, C. M., Baird, A. A., Miller, M. B., & Wolford, G. L. (2011). Neural correlates of interspecies perspective taking in the post-mortem atlantic salmon: an argument for proper multiple comparisons correction. *Journal of Serendipitous and Unexpected Results*, 1: 15. doi: 10.1016/s1053-8119(09)71202-9.

Bergmann, K., Branigan, H. P., & Kopp, S. (2015). Exploring the alignment space – lexical and gestural alignment with real and virtual humans. *Frontiers in ICT*, 2(7): 1-11. doi: 10.3389/fict.2015.00007.

Bergs, A. & Diewald, G. (2009). *Contexts and Constructions*. Amsterdam: John Benjamins.

Berreby, D. (2012). Obese? Smoker? No retirement savings? Perhaps it's because of the language you speak. *Big Think*. February 5, 2012. <http://bigthink.com/ideas/42306>

Berwick, R. C., & Chomsky, N. (2016). *Why only us: Language and Evolution*. Cambridge, MA: MIT Press.

Berwick, R. C., Pietroski, P., Yankama, B., & Chomsky, N. (2011). Poverty of the Stimulus Revisited. *Cognitive Science*, 35: 1207-1242.

Beuls, K., & Steels, L. (2013). Agent-based models of strategies for the emergence and evolution of grammatical agreement. *PLoS ONE*, 8(3): e58960. doi: 10.1371/journal.pone.0058960.

Bibiko, H. J. (2009). *Program for creating nexus files from classificatory language labels (version 1.0)*. Online <http://email.eva.mpg.de/wichmann/software.htm>. Accessed 18/04/2013.

Bickel, B., & Nichols, J. (2011). Inectional synthesis of the verb. In M. S. Dryer, M. Haspelmath, (Eds.), *The World Atlas of Language Structures Online*, Munich: Max Planck Digital Library. Accessed 18/04/2013. <http://wals.info/chapter/22>.

Bivens, J., & Irons, J. (2010). Government debt and economic growth. *Economic policy institute Report: Budget Taxes and Public Investment*, 26/07/2010. <http://www.epi.org/publication/bp271/>. Accessed 18/04/2013.

Bleys, J., & Steels, L. (2009). Linguistic Selection of Language Strategies - A Case Study for Colour. *ECAL*, 2: 150-157.

Blokpoel, M., et al. (2012). Recipient design in human communication: simple heuristics or perspective taking? *Frontiers in Human Neuroscience*, Special Issue: Towards a neuroscience of social interaction. doi: 10.3389/fnhum.2012.00253.

Bloomfield, L. (1933). *Language*. New York, NY: Holt, Rinehart and Winston.

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

Bowler, T. (2013). Why speaking English can make you poor when you retire. *BBC News: Business*. Posted 23/02/2013, accessed 19/04/2013. <http://www.bbc.co.uk/news/business-21518574>.

Bouckaert R., et al. (2012). Mapping the origins and expansion of the Indo-European language family. *Science*, 337: 957-960. doi: 10.1126/science.1219669.

Boyd, R. & Richerson, P. (1985). *Culture and the evolutionary process*, Chicago: University of Chicago Press.

Branigan, H. P., et al. (2011). The role of beliefs in lexical alignment: evidence from dialogs with humans and computers. *Cognition*, 121: 41-57. doi: 10.1016/j.cognition.2011.05.011.

Bratman, M. (1992). Shared cooperative activity. *The Philosophical Review*, 101: 237-241.

Brennan, S. E., & Clark, H. H. (1996). Conceptual packs and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22: 1482-1493.

Brennan, S. E., Galati, A., & Kuhlen, A. K. (2010). Two Minds, One Dialog: Coordinating Speaking and Understanding. In B. H. Ross (Ed.), *The Psychology of Learning and Motivation*, 53: 301-344. Burlington: Elsevier.

Brighton, H., Kirby, S., & Smith, K. (2005). Cultural selection for learnability: Three principles underlying the view that language adapts to be learnable. In M. Tallerman (Ed.), *Language Origins: Perspectives on Evolution*, chapter 13. Oxford: Oxford University Press.

Brinton, L. J., & Traugott, E. C. (2005). *Lexicalization and Language Change*, Cambridge: Cambridge University Press.

Brown, P. M., & Dell, G. S. (1987). Adapting production to comprehension: The explicit mention of instruments. *Cognitive Psychology*, 19: 441-472.

Bussmann, H. (1996). *Routledge Dictionary of Language and Linguistics*, London: Routledge.

Button, K. S., et al. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14: 365-376. doi: 10.1038/nrn3475.

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

Bybee, J. (2002). Sequentiality as the basis of constituent structure. In T. Givon & B. Malle (Eds.), *The evolution of language from pre-language*: 109-132. Amsterdam: John Benjamins.

Bybee, J. (2003). Mechanisms of change in grammaticization: the role of frequency. In B. D. Joseph and R. D. Janda (Eds.), *The handbook of historical linguistics*: 602-23. Oxford: Blackwell.

Bybee, J. (2010). *Language, Usage and Cognition*. Cambridge: Cambridge University Press.

Caldecott, J., Jenkins, M., Johnson, T., & Groombridge, B. (1994). Priorities for conserving global species richness and endemism. In N. M. Collins, (Ed.), *World Conservation Monitoring Centre, Biodiversity Series No. 3*: 17. Cambridge: World Conservation Press.

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

Caspi, A., et al. (2003). Influence of Life Stress on Depression: Moderation by a Polymorphism in the 5-HTTLPR Gene. *Science*, 301: 386-389. doi: 10.1126/science.1083968.

Central Intelligence Agency (2013). The world factbook: Roadways. accessed 18/04/2013. <https://www.cia.gov/library/publications/the-world-factbook/rankorder/2085rank.html>.

Central Intelligence Agency. (2013). The world factbook: Net migration rate. accessed 18/04/2013. URL <https://www.cia.gov/library/publications/the-world-factbook/rankorder/2112rank.html>.

Chafe, W. (1976). Givenness, Contrastiveness, Definiteness, Subjects, Topics and Point of View. In C. N. Li (Ed.), *Subject and Topic*: 25-56. New York: Academic Press.

Channell, J. (1994). *Vague Language*. Oxford: Oxford University Press.

Chater, N., & Christiansen, M. (2010). Language Acquisition Meets Language Evolution. *Cognitive Science*, 34: 1131-1157.

Chater, N., & Vitanyi, P. (2003). Simplicity: A unifying principle in cognitive science. *Trends in Cognitive Science*, 7: 19-22.

Chiao, J. Y., & Blizinsky, K. D. (2010). Culture-gene coevolution of individualism-collectivism and the serotonin transporter gene. *Proceedings of the Royal Society B: Biological Sciences*: 277.

Chen, M. K. (2012). Could your language affect your ability to save money? *TEDGlobal 2012*, Online: http://www.ted.com/talks/keith_chen_could_your_language_affect_your_ability_to_save_money.html. Accessed 18/04/2013.

Chen, M. K. (2013). The effect of language on economic behavior: Evidence from savings rates, health behaviors, and retirement assets. *American Economic Review*, 103: 690-731. doi: 10.1257/aer.103.2.690.

Chomsky, N. (1957). *Syntactic Structures*. New York, NY: Mouton de Gruyter.

Chomsky, N. (1988). *Language and Problems of Knowledge: The Managua Lectures*. London: MIT Press.

Chomsky, N. (2002). *On Nature and Language*. Cambridge: Cambridge University Press.

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

Christiansen, M. & Kirby, S. (2003). *Language Evolution*. Oxford: Oxford University Press.

Claesson, K. (1994). A phonological outline of mataco-noctenes. *International Journal of American Linguistics*, 60: 1-38. doi: 10.1086/466216.

Clahsen, H., Felser, C., Neubauer, K., Sato, M., & Silva, R. (2010). Morphological structure in native and nonnative language processing. *Language Learning*, 60: 21-43. doi: 10.1111/j.1467-9922.2009.00550.x.

Clark, A. (2015). Radical Predictive Processing. *The Southern Journal of Philosophy*, 53: 3-27. doi: 10.1111/sjp.12120.

Clark, H. H. (1992). *Arenas of Language Use*. Chicago: University of Chicago Press.

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

Clark, H. H., & Carlson, T. B. (1992). Context for Comprehension. In H. H. Clark, *Arenas of Language Use*, 60-77. Chicago: University of Chicago Press.

Clark, H. H. & Marshall, C. R. (1981). Definite reference and mutual knowledge. In A. K. Joshi, B. L. Webber, & I. A. Sag (Eds.), *Elements of Discourse Understanding*: 1063. Cambridge: Cambridge University Press.

Clark, H. H., & Murphy, G. L. (1982). Audience design in meaning and reference. *Advances in Psychology*. 9: 287-299. doi: 10.1016/S0166-4115(09)60059-5.

CMS collaboration and others. (2012). Observation of a new boson with a mass near 125 gev. *CMS physics analysis summary CMSPAS-HIG-12 20*.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, second edition. Mahwah, NJ: Lawrence Erlbaum.

Cohen Priva, U. (2008). Using information content to predict phone deletion. In N. Abner & J. Bishop (Eds.), *Proceedings of the 27th west coast conference on formal linguistics*: 90-98.

Collard, M., Shennan, S. J., & Tehrani, J. J. (2006). Branching, blending, and the evolution of cultural similarities and differences among hu-

man populations. *Evolution and Human Behavior*, 27: 169-184. doi: 10.1016/j.evolhumbehav.2005.07.003.

Congdon, P. (2006). *Bayesian models for categorical data*. New York, NY: John Wiley & Sons Ltd.

Cornish, H. (2010). Investigating how cultural transmission leads to the appearance of design without a designer in human communication systems. *Interaction Studies*, 11(1): 112-137.

Cornish, H., Tamariz, M. & Kirby, S. (2009). Complex adaptive systems and the origins of adaptive structure: what experiments can tell us. Special issue on Language as a Complex Adaptive System. *Language Learning*, 59(s1): 187-205.

Croft, W. (2000). *Explaining Language Change: An evolutionary approach*. Harlow: Longman.

Crop Protection Compendium. (2008). *Acacia confusa*. Accessed 18-04-2011. <http://www.cabi.org>.

Cruse, D. A. (2011). *Meaning in language: An introduction to semantics and pragmatics*. Oxford, UK: Oxford University Press.

Culbertson, J., & Adger, D. (2014). Language learners privilege structured meaning over surface frequency. *PNAS*, 111(16): 5842-5847.

Culbertson, J., & Kirby, S. (2016). Simplicity and specificity in language: Domain general biases have domain specific effects. *Frontiers in Psychology*, 6(1964). doi: 10.3389/fpsyg.2015.01964.

Culbertson, J., Smolensky, P., & Wilson, C. (2013). Cognitive biases, linguistic universals, and constraint-based grammar learning. *Topics in Cognitive Science*, 5(3):392-424.

Cutting, J., (2007). *Vague Language Explored*. New York, NY: Palgrave Macmillan.

Dahl, Ö. (2004). *The growth and maintenance of linguistic complexity*. John Benjamins Publishing Company.

Dahl, Ö. (2013). Stuck in the futureless zone. *Diversity Linguistics comment*. Posted 03/09/2013. Accessed 18/04/2013. <http://dlchypotheses.org/360>.

Dale, R., & Lupyan, G. (2012). Understanding the origins of morphological diversity: The linguistic niche hypothesis. *Advances in Complex Systems*, 15: 1150017.

Dale, R., & Reiter, E. (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19: 233-263.

Darwin, C. (1871). *The Descent of Man and Selection in Relation to Sex*. London: John Murray.

Dediu, D. (2008). The role of genetic biases in shaping the correlations between languages and genes. *Journal of theoretical biology*, 254: 400-407. doi: 10.1016/j.jtbi.2008.05.028.

Dediu, D. (2013). Genes: Interactions with Language on Three Levels – Inter-individual Variation, Historical Correlations and Genetic Biasing. In P. M.

Binder & K. Smith. (Eds), *The Language Phenomenon*, Berlin, Germany: Springer-Verlag.

Dediu, D., & Ladd, D. (2007). Linguistic tone is related to the population frequency of the adaptive haplogroups of two brain size genes, ASPM and Microcephalin. *Proceedings of the National Academy of Sciences*, 104: 10944. doi: 10.1073/pnas.0610848104.

De Ruiter, J., et al. (2010). Exploring the cognitive infrastructure of communication. *Interaction Studies*, 11: 51-77.

de Saussure, F. (1916). *Course in general linguistics*. New York, NY: McGraw-Hill

Desideri, G., et al. (2012). Benefits in cognitive function, blood pressure, and insulin resistance through cocoa flavanol consumption in elderly subjects with mild cognitive impairment novelty and significance the cocoa, cognition, and ageing (cocoa) study. *Hypertension*, 60: 794-801. doi: 10.1161/hypertension-aha.112.193060.

Diesendruck, G., & Boom, P. (2003). How specific is the Shape Bias? *Child Development*, 74(1): 168-178.

Dingemans, M., Torreira, F., Enfield, N. J. (2013). Is Huh? a Universal Word? Conversational Infrastructure and the Convergent Evolution of Linguistic Items. *PLoS ONE*, 8(11): e78273. doi: 10.1371/journal.pone.0078273.

Donohue, M., & Nichols, J. (2011). Does phoneme inventory size correlate with population size? *Linguistic Typology*, 15: 161-170. doi: 10.1515/lity.2011.011.

Dryer, M. & Haspelmath, M. (2013). *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Du Bois, J. W. (1987). The Discourse Basis of Ergativity. *Language*, 63(4): 805-855.

Dunn, M., Greenhill, S. J., Levinson, S. C., & Gray, R. D. (2011). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, 473: 79-82.

Duranti, A. & Goodwin, A. (1992). *Rethinking Context: Language as an interactive phenomenon*. Cambridge: Cambridge University Press.

Eisenberg, D. T., & Hayes, M. G. (2011). Testing the null hypothesis: comments on culture-gene coevolution of individualismcollectivism and the serotonin transporter gene. *Proceedings of the Royal Society B: Biological Sciences*, 278: 329-332. doi: 10.1098/rspb.2010.0714.

Elman, J. L. (2008). The shape bias: an important piece in a bigger puzzle. *Developmental Science*, 11(2): 219-222.

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science*, 33(4): 547-582.

Ember, C., & Ember, M. (2007). Climate, econiche, and sexuality: Influences on sonority in language. *American Anthropologist*, 109: 180-185. doi: 10.1525/aa.2007.109.1.180.

Engelkamp, J., Zimmer, H. D. & Mohr, G. (1990). Differential memory effects of concrete nouns and action verbs. *Zeitschrift für Psychologie*, 198: 189-216.

- Estill, R., & Kemper, S. (1982). Interpreting idioms. *Journal of Psycholinguistic Research*, 9: 559-568.
- Evans, N. & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32(5): 429-492.
- Evans, V. (2005). The meaning of time: Polysemy, the lexicon and conceptual structure. *Journal of Linguistics*, 41: 33-75.
- Evans, V. & Green, M. (2006). *Cognitive Linguistics: An Introduction*. Edinburgh: Edinburgh University Press.
- Everett, C., Blasi, D. E., & Roberts, S. G. (2015). Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots. *Proceedings of the National Academy of Sciences of the United States of America*, 112: 1322-1327. doi:10.1073/pnas.1417413112.
- Faber, P. & León-Araúz, P. (2016). Specialized knowledge representation and the parameterization of context. *Frontiers in Psychology*, 7. doi: 10.3389/fpsyg.2016.00196.
- Fanelli, D. (2010). Positive results increase down the hierarchy of the sciences. *PloS ONE*, 5: e10068. doi: 10.1371/journal.pone.0010068.
- Fauconnier, G. (1985). *Mental Spaces*. Cambridge, MA: MIT Press.
- Fay, N., & Ellison, T. M. (2013). The Cultural Evolution of Human Communication Systems in Different Sized Populations: Usability Trumps Learnability. *PLoS ONE*, 8(8): e71781. doi:10.1371/journal.pone.0071781.

Fay, N., Garrod, S., & Roberts, L. (2008). The fitness and functionality of culturally evolved communication systems. *Philosophical Transactions of the Royal Society London B Biological Sciences*, 363(1509): 3553-3561.

Fedzechkina, M., Jaeger, T. F., & Newport, E. L. (2012). Language learners restructure their input to facilitate efficient communication. *PNAS*, 109: 17897-17902.

Fellman, B. (2012). Speaking and saving. *Yale Alumni Magazine*, January 1st Edition.

Ferreira, V., Slevc, L., & Rogers, E. (2005). How do speakers avoid ambiguous linguistic expressions? *Cognition*, 96(3): 263-284.

Fery, C. & Krifka, M. (2008). Information structure. Notional distinctions, ways of expression. In P. V. Sterkenburg (Ed.), *Unity and diversity of languages*: 123-136, Amsterdam: John Benjamins.

Fetzer, A. (2004). *Recontextualizing Context*. New York, NY: Benjamins.

Fillmore, C. (1985). Frames and the Semantics of Understanding. *Quaderni Di Semantica*, 6(2): 222-254.

Finkbeiner, R., Meibauer, J. & Schumacher, P. B. (2012). *What is a Context? Linguistic approaches and challenges*. Amsterdam: John Benjamins.

Fought, J. G., Munroe, R. L., Fought, C. R., & Good, E. M. (2004). Sonority and climate in a world sample of languages: Findings and prospects. *Cross-cultural research*, 38: 27-51. doi: 10.1177/1069397103259439.

- Fortson, B. W. (2004). An approach to semantic change. In B. D. Joseph & R. D. Janda (Eds.), *The Handbook of Historical Linguistics*: 648-666.
- Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336: 998.
- Frank, M. C., & Goodman, N. D. (2014). Inferring word meanings by assuming that speakers are informative. *Cognitive Psychology*, 75: 80-96.
- Franke, M. (2013). Game Theoretic Pragmatics. *Philosophy Compass*, 8(3): 269-284.
- Franke, M., & Degen, J. (2016). Reasoning in reference games: individual- vs. population-level probabilistic modeling. *PLoS ONE*, 11(5): 1-25.
- Franke, M., & Jäger, G. (2016). Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft*, 35(1): 344.
- Frege, G. (1884). *Die Grundlagen der Arithmetik: eine logisch-mathematische Untersuchung ber den Begriff der Zahl*, Breslau: W. Koebner. Translated as *The Foundations of Arithmetic: A logico-mathematical enquiry into the concept of number*, by J.L. Austin, Oxford: Blackwell, second revised edition, 1953.
- Freyd, J. (1983). Shareability: The social psychology of epistemology. *Cognitive Science*, 7(3): 191-210.
- Fritz, S. A., & Purvis, A. (2010). Selectivity in mammalian extinction risk and threat types: a new measure of phylogenetic signal strength in binary traits. *Conservation Biology*, 24: 1042-1051. doi: 10.1111/j.1523-1739.2010.01455.x.

- Galantucci, B. (2005). An experimental study of the emergence of human communication systems. *Cognitive Science*, 29(5): 737-767.
- Galantucci, B., Garrod, S., & Roberts, G. (2012). Experimental Semiotics. *Language and Linguistics Compass*, 6: 477-493. doi: 10.1002/lnc.
- Galantucci, B., Kroos, C., & Rhodes, T. (2010). The effects of rapidity of fading on communication systems. *Interaction Studies*, 11: 100-111.
- Gärdenfors, P. (2000). *Conceptual Spaces: the geometry of thought*. Cambridge, MA: MIT Press.
- Garrod, S., & Galantucci, B. (2011). Experimental Semiotics: A Review. *Frontiers in Human Neuroscience*, 168(6): 11.
- 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.
- Geeraerts, D. (1993). Vagueness's puzzles, polysemy's vagaries. *Cognitive Linguistics*, 4(3): 223-272.
- Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66: 8-38.
- Geraint, S. (2011). Language evolution and the acacia tree. *SpecGram CLXII*.
- Gibbs, R., & Gonzales, G. P. (1985). Syntactic frozenness in processing and remembering idioms. *Cognition*, 20(3): 243-259.
- Gibson, E. J. & Pick, A. D. (2000). *An ecological approach to perceptual learning and development*. New York: Oxford University Press.

Gigerenzer, G. & Brighton, H. (2009). Homo heuristicus: why biased minds make better inferences. *Topics in Cognitive Science*, 1: 107-43.

Gigerenzer, G. & Gaissmaier, W. (2011). Heuristic Decision Making. *Annual Review of Psychology*, 62: 451-482.

Gil, D. (2005). Word order without syntactic categories: How Riau Indonesian does it. In A. Carnie, H. Harley, & S. A. Dooley (Eds.), *Verb First: On the Syntax of Verb-initial Languages*: 243-264. Amsterdam: John Benjamins.

Giles, H., Coupland, N., & Coupland, J. (1991). *Contexts of Accommodation*. Cambridge: Cambridge University Press.

Givon, T. (2005). *Context as Other Minds: The Pragmatics of Sociality, Cognition and Communication*. Amsterdam: John Benjamins.

Goffman, E. (1974). *Frame analysis: An essay on the organization of experience*. Cambridge, MA: Harvard University Press.

Gordon, R. (2005). *Ethnologue: Languages of the World, 15th Edition*. SIL International.

Gould, S. J. (1987). *The limits of adaptation: is language a spandrel of the human brain?*. Paper presented to to the Cognitive Science Seminar, Center for Cognitive Science, MIT, Cambridge, MA.

Graur, D., et al. (2013). On the immortality of television sets: function in the human genome according to the evolution-free gospel of ENCODE. *Genome Biology and Evolution*. doi: 10.1093/gbe/evt028.

Gray, R. D., Drummond, A. J., & Greenhill, S. J. (2009). Language phylogenies reveal expansion pulses and pauses in pacific settlement. *Science*, 323: 479-483. doi: 10.1126/science.1166858.

Gray, R. D., & Jordan, F. M. (2000). Language trees support the express-train sequence of Austronesian expansion. *Nature*, 405: 1052-1055.

Greenberg, J. (1963). Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements. In Greenberg, Joseph H. (ed.), *Universals of Human Language*: 73-113. Cambridge, Mass: MIT Press.

Grice, H. P. (1957). Meaning. *Philosophical Review*, 66: 377-388.

Griffiths, T. L., & Kalish, M. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science*, 31: 441-480.

Grodner, D. & Sedivy, J. (2011). The effects of speaker-specific information on pragmatic inferences. In N. Pearlmuter & E. Gibson (Eds). *The processing and acquisition of reference*: 239-272. Cambridge, MA: MIT Press.

Harmon, L., et al. (2009). *geiger: Analysis of evolutionary diversification*. <http://CRAN.R-project.org/package=geiger>. R package version 1.31. Accessed 18/04/2013.

Harrell Jr., F. E., et al. (2014). *Hmisc: Harrell Miscellaneous*. <http://CRAN.R-project.org/package=Hmisc>. R package version 3.14-5.

Harris, A., & Campbell, L. (1995). *Historical syntax in cross-linguistic perspective*. Cambridge, UK: Cambridge University Press.

Haspelmath, M. (2004). On directionality in language change with particular reference to grammaticalization. In O. Fischer, M. Norde & H. Perridon (Eds.), *Up and down the cline: The nature of grammaticalization (Typological Studies in Language, 59)*: 17-44. Amsterdam: John Benjamins.

Hay, J. (2001). Lexical frequency in morphology: is everything relative? *Linguistics*, 39: 1041-1070.

Hay, J. (2002). From speech perception to morphology: affix-ordering revisited. *Language*, 78: 527-555.

Hay, J., & Bauer, L. (2007). Phoneme inventory size and population size. *Language*, 2: 388-400. doi: 10.1353/lan.2007.0071.

Hayashi, M., Raymond, G., & Sidnell, J. (2013). *Conversational Repair and Human Understanding*. Cambridge: Cambridge University Press.

Haywood, S. L., Pickering, M. J., & Branigan, H. P. (2005). Do speakers avoid ambiguities during dialogue? *Psychological Science*, 16(5): 362-6.

Heine, B., Claudi, U., & Hünnemeyer, F. (1991). *Grammaticalization: a conceptual framework*. Chicago, IL: University of Chicago Press.

Heller, D., Skovbrot, K., & Tanenhaus, M. K. (2009). Experimental evidence for speakers' sensitivity to common vs. privileged ground in the production of names. *PRE-CogSci Workshop on the Production of Referring Expressions*. Amsterdam, Netherlands.

Herndon, T., Ash, M., & Pollin, R. (2013). Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Political Economy Research Institute Working paper series*, 322.

- Hinton, A. (2014). *Understanding Context: Environment, Language, and Information Architecture*. Newton, MA: O'Reilly Media.
- Hockett, C. (1960). The Origin of Speech. *Scientific American*, 203: 88-111.
- Hoefer, S. H. (2009). *Modelling the role of pragmatic plasticity in the evolution of linguistic communication*. PhD Thesis, University of Edinburgh.
- Hopper, P. J., & Traugott, E. C. (2003). *Grammaticalization*. 2nd Edition. Cambridge: Cambridge University Press.
- Horn, L. R. (1993). Economy and redundancy in a dualistic model of natural language. In S. Shore & M. Vilkuna (Eds.), *SKY 1993: 1993 Yearbook of the Linguistic Association of Finland*: 33-72.
- Horton, W. S., & Keysar, B. (1996). When Do Speakers Take into Account Common Ground? *Cognition*, 59(1): 91-117. doi:10.1016/0010-0277(96)81418-1.
- Horton, W. S., & Gerrig, R. J. (2005). The impact of memory demands on audience design during language production. *Cognition*, 96: 127-142.
- Hurford, J. R. (2003). The language mosaic and its evolution. *Studies in the Evolution of Language*, 3: 38-57.
- Hurford, J. R. (2011). *The Origins of Grammar: Language in the Light of Evolution II*. Oxford: Oxford University Press.
- Hurlburt, R. T., & Knapp, T. J. (2006). Münsterberg in 1898, Not Allport in 1937, Introduced the Terms Idiographic and Nomothetic to American Psychology. *Theory & Psychology*, 16(2): 287-293.

Husten, L. (2012). Chocolate and nobel prizes linked in study. *Forbes: Pharma & Healthcare*. <http://www.forbes.com/sites/larryhusten/2012/10/10/chocolate-and-nobel-prizes-linked-in-study/>. Accessed 18/04/2013/

Hymes, D. (1987). Communicative competence. In U. Ammon, N. Dittmar & K. J. Mattheier (Eds.), *Sociolinguistics: An international Handbook of the Science of Language and Society*. 219-229. Berlin: Walter de Gruyter.

Irvine, L., Roberts, S. G., & Kirby, S. (2013). A robustness approach to theory building: A case study of language evolution. In M. Knauff et al., (Eds.), *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*: 2641-2619.

Isaacs, E. A., & Clark, H. H. (1987). References in conversation between experts and novices. *Journal of Experimental Psychology*, 116: 26-37. doi: 10.1037/0096-3445.116.1.26.

Jaeger, T. F., Graff, P., Croft, W., & Pontillo, D. (2011). Mixed effect models for genetic and areal dependencies in linguistic typology. *Linguistic Typology*, 15: 281-319. doi: 10.1515/lity.2011.021.

Joelving, F. (2012). Eat chocolate, win the nobel prize? *Reuters US*, Online 10/10/20 <http://www.reuters.com/article/2012/10/10/us-chocolate-nobels-idUSBRE8991SS20121010>. Accessed 18/04/2013.

Jones, S. S., Smith, L. B., & Landau, B. (1991). Object properties and knowledge in early lexical learning. *Child Development*, 62(3): 499-516.

Juba, B., Kalai, A., Khanna, S., & Sudan, M. (2011). Compression without a common prior: An information-theoretic justification for ambiguity in

language. *Proceedings of the Second Symposium on Innovations in Computer Science*.

Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H., & Bühlmann, P. (2012). Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47: 126.

Kamide, Y., Altmann, G., & Haywood, S. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, 49(1): 133-156.

Kay, P. (1977). Language evolution and speech style. In B. G. Blount & M. Sanches (Eds.), *Sociocultural dimensions of language change*: 21-33. New York, NY: Academic Press.

Keating, J. E. (2012). Tomorrow, we save. *Foreign Policy* 01/10/2012 Online http://www.foreignpolicy.com/articles/2012/08/13/tomorrow_we_save.html. Accessed 18/04/2013.

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, 336: 1049-1054.

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, 11: 3238.

Kiparsky, P. (2012). Grammaticalization as optimization. In D. Jonas, J. Whitman & A. Garrett (Eds.), *Grammatical Change: Origins, Nature, Outcomes*. Oxford: Oxford University Press.

Kirby, S. (1999). *Function, Selection and Innateness: the Emergence of Language Universals*. Oxford: Oxford University Press.

Kirby, S. (2012). Language is an Adaptive System. The Role of Cultural Evolution in the Origins of Structure. In M. Tallerman & K. R. Gibson (Eds.), *The Oxford Handbook of Language Evolution*: 589-604. Oxford: Oxford University Press.

Kirby, S. (2013). The Evolution of Linguistic Replicators. In K. Smith & P. Binder (Eds.), *The Language Phenomenon*. New York, NY: Springer.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *PNAS*, 105(31): 10681-10686.

Kirby, S., Dowman, M., & Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *PNAS*, 104(12): 5241-5245. doi: 10.1073/pnas.0608222104

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current opinions in neurobiology*, 28: 108-114.

Kirby, S. & Hurford, J. R. (2002). The emergence of Linguistic Structure: An overview of the Iterated Learning model. In A. Cangelosi and D. Parisi (Eds.), *Simulating the Evolution of Language*. New York, NY: Springer.

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

Konopka, A. E. & Brown-Schmidt, S. (2014). Message encoding. In V. Ferreira, M. Goldrick, and M. Miozzo (Eds.), *The Oxford handbook of language production*: 3-20. New York: Oxford University Press.

Koolen, R., Goudbeek, M., & Krahmer, E. (2013). The effect of scene variation on the redundant use of color in definite references. *Cognitive Science*, 37: 395-411.

Kuteva, T. (2001). *Auxiliation: An enquiry into the nature of grammaticalization*. Oxford: Oxford University Press.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2016). *lmerTest: Tests in Linear Mixed Effects Models*. R package version 2.0-30. <http://CRAN.R-project.org/package=lmerTest>.

Labov, W. (2007). Transmission and Diffusion. *Language*, 83: 334-387.

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

Ladd, D. R., et al. (2013). Patterns of individual differences in the perception of missing-fundamental tones. *Journal of Experimental Psychology: Human Perception and Performance*, PMID: 23398251.

Langacker, R. (1987). The form and meaning of the English auxiliary. *Language*, 54: 853-82.

Lass, R. (1997). *Historical linguistics and language change*. Cambridge: Cambridge University Press.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Levinson, S. C. (2000). *Presumptive meanings: the theory of generalised conversational implicature*. Cambridge, MA: MIT Press.

Levinson, S. C. (2006). On the human interaction engine. In N. Enfield, & S. Levinson, (Eds.), *Roots of Human Sociality: Culture, Cognition and Human Interaction*: 39-69. Oxford: Berg.

Levinson, S. C. (2016). Turn-taking in human communication, origins, and implications for language processing. *Trends in Cognitive Sciences*, 20(1): 6-14. doi: 10.1016/j.tics.2015.10.010.

Levinson S. C., & Gray, R. (2012). Tools from evolutionary biology shed new light on the diversification of languages. *Trends in Cognitive Sciences*, 16: 167173. doi: 10.1016/j.tics.2012.01.007.

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

Lewis, M. L., & Frank, M. C. (2016). The length of words reflects their conceptual complexity. *Cognition*, 153: 182-195.

Lewis, M. L., Sugarman, E., & Frank, M. C. (2014). The structure of the lexicon reflects principles of communication. In P. Bello, M. Guarini, M. McShane & B. Scassellati (Eds.), *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Lewis, M. P. (2009). *Statistical Summaries*, Dallas, Texas: SIL International, *Ethnologue: Languages of the World, Sixteenth edition*.

Liberman, M. (2012). Cultural diffusion and the whorfian hypothesis. *Language Log*, Posted February 12, 2012, accessed May 21, 2012. <http://language.log1d.cupenn.edu/n11/?p=3764>.

Lightfoot, D. W. (1979). *Principles of diachronic syntax*. Cambridge: Cambridge University Press.

Lindblom, B., MacNeilage, P., & Studdert-Kennedy, M. (1984). Self-organizing processes and the explanation of phonological universals. In B. Butterworth, B. Comrie, & O. Dahl (Eds.), *Explanations for language universals*: 181-203. New York, NY: Mouton.

Lloyd, K. & Leslie, D. S. (2013). Context-dependent decision-making: a simple Bayesian model. *Journal of the Royal Society Interface*, 10(82). doi: 10.1098/rsif.2013.0069.

Lupyan, G., & Clark, A. (2015). Words and the World: Predictive coding and the Language-Perception-Cognition Interface. *Current directions in Psychological Science*, 24(4): 279-284.

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

Lynn, R., & Vanhanen, T. (2002). *IQ and the wealth of nations*. London: Praeger Publishers.

Lyons, J. (1977). *Semantics*. Cambridge: Cambridge University Press.

Maathuis, M. H., Colombo, D., Kalisch, M., & Bühlmann, P. (2010). Predicting causal effects in large-scale systems from observational data. *Nature Methods*, 7: 247-248. doi: 10.1038/nmeth0410-247.

Mace, R., & Jordan, F. M. (2011). Macro-evolutionary studies of cultural diversity: A review of empirical studies of cultural transmission and cultural adaptation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366: 402-411. doi: 10.1098/rstb.2010.0238.

Maddieson, I. (2011). Consonant inventories. In M. S. Dryer, M. Haspelmath, (Eds.), *The World Atlas of Language Structures Online*. Accessed 18/04/2013, Munich: Max Planck Digital Library. <http://wals.info/feature/1A>.

Maddieson, I. (2011). Tone. In M. S. Dryer, M. Haspelmath, (Eds.), *The World Atlas of Language Structures Online*. Accessed 18/04/2013, Munich: Max Planck Digital Library. <http://wals.info/chapter/13>.

Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/Information theory: Speakers choose shorter words in predictive contexts. *Cognition*, 126(2): 313-318.

Marcus, G. (2013). Steamrolled by big data. *The New Yorker Elements Blog*. Posted 03/04/2013, accessed 19/04/2013. Online: <http://www.newyorker.com/online/blogs/elements/2013/04/steamrolled-by-big-data.html>.

Markman, E., & Wachtel, G. (1988). Childrens use of mutual exclusivity to constrain the meanings of words. *Cognitive psychology*, 157: 121-157.

Martins, E. P., & Hansen, T. F. (1997). Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *American Naturalist*. 646-667.

Mesoudi, A. (2011). *Cultural Evolution: How Darwinian theory can explain human culture and synthesize the social sciences*. Chicago, IL: University of Chicago Press.

Messerli, F. H. (2012). Chocolate consumption, cognitive function, and nobel laureates. *New England Journal of Medicine*, 367: 1562-1564. doi: 10.1056/nejmon1211064.

Mihatsch, W. (2009). Nouns are THINGS: Evidence for a grammatical metaphor? In K. U. Panther, L. L. Thornburg & A. Barcelona. *Metonymy and Metaphor in Grammar*. Amsterdam, Netherlands: John Benjamins.

Miller, G. A. (1951). *Language and Communication*. New York, NY: McGraw-Hill Book Company.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 63: 81-97.

Millikan, R. G. (1998). Language conventions made simple. *The Journal of Philosophy*, 95(4): 161-180.

Milosavljevic, M., Navalpakkam, V., Koch, C., & Rangel, A. (2012). Relative visual saliency differences induce sizable bias in consumer choice. *Journal of Consumer Psychology*, 22(1): 67-74.

Minsky, M. A. (1975). A framework for representing knowledge. In M. Winston (Ed.), *The psychology of computer vision*. Boston, MA: MIT Press.

Mitchell, T., Hulme, M., & New, M. (2002). Climate data for political areas. *Area*, 34: 109112. doi: 10.1111/1475-4762.00062.

Moran, S. (2012). *Phonetics Information Base and Lexicon*. Ph.D. thesis, University of Washington.

- Moran, S., McCloy, D., & Wright, R. (2012). Revisiting population size vs. phoneme inventory size. *Language*, 88: 877-893. doi: 10.1353/lan.2012.0087.
- Moreno, M., & Baggio, G. (2015). Role Asymmetry and Code Transmission in Signaling Games: An Experimental and Computational Investigation. *Cognitive Science*, 39(5): 918-943. doi: 10.1111/cogs.12191.
- Mufwene, S. (2001). *The Ecology of Language Evolution*. Cambridge, UK: Cambridge University Press.
- Murdock, G., & White, D. (1969). Standard cross-cultural sample. *Ethnology*, 8: 329-369. doi: 10.2307/3772907.
- Murray, E. J. (1965). *Sleep, dreams, and arousal*. New York, NY: Appleton-Century-Crofts.
- Nadig, A. S. & Sedivy, J. C. (2002). Evidence of perspective-taking constraints in childrens on-line reference resolution. *Psychological Science*, 13: 329-336.
- Naroll, R. (1961). Two solutions to Galton's problem. *Philosophy of Science*, 28: 15-39. doi: 10.1086/287778.
- Nerlich, B., & Clarke, D. D. (1992). Outline of a model for semantic change. In G. Kellermann & M. D. Morrissey (Eds.), *Diachrony within Synchrony: Language History and Cognition*: 125-144. Frankfurt am Main: Peter Lang.
- Nettle, D. (1998). Explaining global patterns of language diversity. *Journal of anthropological archaeology*, 17: 354-374. doi: 10.1006/jaar.1998.0328.
- Nettle, D. (1999). *Linguistic Diversity*. New York, NY: Oxford University Press.

Nettle, D. (2007). Language and genes: A new perspective on the origins of human cultural diversity. *Proceedings of the National Academy of Sciences of the USA*, 104: 10755-6.

Nettle, D. (2009). Ecological influences on human behavioural diversity: a review of recent findings. *Trends in ecology & evolution*, 24: 618-624. doi: 10.1016/j.tree.2009.05.013.

Nettle, D. (2012). Social scale and structural complexity in human languages. *Philosophical Transactions of the Royal Society B*, 367: 1829-36.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: MIT Press.

Nowak, I., & Baggio, G. (2016). The emergence of word order and morphology in compositional languages via multigenerational signaling games. *Journal of Language Evolution*, Advance Access: 1-14. doi: 10.1093/jole/lzw007.

Nunan, D. (1992). *Research Methods in Language Learning*. Cambridge: Cambridge University Press.

Nunn C. L.(2013). *The AnthroTree website*. Online: <http://nunn.rc.fas.harvard.edu/groups/pica/>. Accessed 18/04/2013.

Nunn, C. L., Mulder, M. B., & Langley, S. (2006). Comparative methods for studying cultural trait evolution: A simulation study. *Cross-Cultural Research*, 40: 177-209. doi: 10.1177/1069397105283401.

Olson, D. R. (1970). Language and thought: Aspects of a cognitive theory of semantics. *Psychological Review*, 77: 257-273.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251). doi: 10.1126/science.aac4716.

Orme, D., et al. (2012). *CAPER: Comparative Analyses of Phylogenetics and Evolution in R*. <http://CRAN.R-project.org/package=caper>.R. Package version 0.5. Accessed 18/04/2013.

Pagel, M. D. (1992). A method for the analysis of comparative data. *Journal of theoretical Biology*, 156: 431-442. doi: 10.1016/s0022-5193(05)80637-x.

Pagel, M. D. (1999). Inferring the historical patterns of biological evolution. *Nature*, 401: 877-884. doi: 10.1038/44766.

Paradis, E., & Claude, J. (2002). Analysis of comparative data using generalized estimating equations. *Journal of Theoretical Biology*, 218: 175-185. doi: 10.1006/jtbi.2002.3066.

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20: 289-290. doi: 10.1093/bioinformatics/btg412.

Parikh, P. (2001). *The Use of Language*. Stanford University: CSLI Publications.

Partee, B. (1984). Compositionality. In F. Landman and F. Veltman (Eds.), *Varieties of Formal Semantics*: 281-312. Reprinted in B. H. Partee. (2004). *Compositionality in Formal Semantics: Selected Papers by Barbara H. Partee*: 153-181. Oxford, UK: Blackwell Publishing.

Paul, H. (1897). *Deutsches Wörterbuch*. Halle: Niemeyer.

Paul, H. (1890). *Principles of the History of Language*. Translated by H. A. Strong from the 2nd German Edition. Reprinted College Park, MD: McGrath (1970).

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, 27: 89-110.

Perkins, R. D. (1992). *Deixis, grammar, and culture*. Amsterdam: John Benjamins Publishing Company.

Perfors, A. & Navarro, D. J. (2014). Language evolution can be shaped by the structure of the world. *Cognitive Science*, 38(4): 775-93.

Pfeifer, N. (2016). Experimental probabilistic pragmatics beyond Bayes' theorem. Commentary on Franke & Jäger: Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft*, 35(1): 89-96.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9): 3526.

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, 122: 280-291.

Pinker, S. (1994). *The language instinct: How the mind creates language*. New York, NY: Harper Collins.

Pinker, S. & Bloom, P. (1990). Natural language and natural selection. *Behavioral and Brain Sciences*, 13: 707-784.

- Pleyer, M., & Winters, J. (2014). Integrating Cognitive Linguistics and language evolution research. *Theoria et Historia Scientiarum*, 11: 19-43.
- Pritchard, C. (2012). Does chocolate make you clever? *BBC News Magazine Online*, 19/11/2012 <http://www.bbc.co.uk/news/magazine-20356613>. Accessed 18/04/2013.
- Pullum, G. K. (2012). Keith Chen, whorfian economist. *Language Log*, Posted February 9, 2012, accessed May 21, 2012 <http://languagelog.ldcupenn.edu/n11/?p=3756>.
- Quasthoff, U. M. (1998). Context. In J. L. Mey & R. E. Asher (Eds.), *Concise Encyclopedia of Pragmatics*: 157-165. Oxford: Elsevier.
- Quine, W. V. O. (1960). *Word and object*. Cambridge, MA: MIT Press.
- R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: <http://www.R-project.org/>.
- Ramscar, M. & Port, R. (2015). Categorization (without categories). In E. Dawbroska & D. Divjak (Eds.), *Handbook of Cognitive Linguistics*. De Gruyter Mouton.
- Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thrope, K. (2010). The effects of Feature-Label-Order and their implications for symbolic learning. *Cognitive Science*, 34: 909-957.
- Reali, F. & Griffiths, T. L. (2009). The evolution of linguistic frequency distributions: Relating regularisation to inductive biases through iterated learning. *Cognition*, 111: 317-328.

- Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, 6: 855-863.
- Ritt, N. (2004). *Selfish sounds and linguistic evolution: A Darwinian approach to language change*. Cambridge: Cambridge University Press.
- Roberts S. G., & Winters, J. (2012). Social structure and language structure: The new nomothetic approach. *Psychology of Language and Communication*, 16: 89-112.
- Roberts, S. G., & Winters, J. (2013). Linguistic Diversity and Traffic Accidents: Lessons from Statistical Studies of Cultural Traits. *PLoS ONE*, 8(8): e70902. doi: 10.1371/journal.pone.0070902.
- Roberts, S. G., Winters, J., & Chen, K. (2015). Future Tense and Economic Decisions: Controlling for Cultural Evolution. *PLoS ONE*, 10(7): e0132145. doi: 10.1371/journal.pone.0132145.
- Rogoff, K., & Reinhart, C. (2010). Growth in a time of debt. *American Economic Review*, 100: 573-78. doi: 10.1257/aer.100.2.573.
- Rohde, H., et al. (2012). Communicating with cost-based implicature: A game-theoretic approach to ambiguity. *Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue*.
- Rosch, E., & Mervis, C. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7(4): 573-605.
- Ross, M. H., & Homer, E. (1976). Galton's problem in cross-national research. *World Politics*, 29: 128. doi: 10.2307/2010045.

Rubio-Fernández, P. (2016). How redundant are redundant colour adjectives? An efficiency-based analysis of color overspecification. *Frontiers in Psychology*. Special issue on Models of Reference.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274(5294): 1926-1928.

Sauce, B., & Matzel, L. D. (2013). The causes of variation in learning and behavior: why individual differences matter. *Frontiers in Psychology*, 4: 395.

Schelling, T. C. (1980). *The strategy of conflict*. Cambridge, MA: Harvard University Press.

Scott-Phillips, T. C. (2015). *Speaking Our Minds: Why human communication is different, and how language evolved to make it special*. Basingstoke: Palgrave Macmillan.

Scott-Phillips, T. C., & Kirby, S. (2010). Language evolution in the laboratory. *Trends in Cognitive Sciences*, 14(9): 411-417.

Scott-Phillips, T. C., Kirby, S., & Ritchie, G. R. S. (2009). Signalling signalhood and the emergence of communication. *Cognition*, 113(2): 226-33.

Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research*, 32: 3-23.

Sedivy, J. C. (2005). Evaluating explanations for referential context effects: Evidence for Gricean Mechanisms in Online Language Interpretation. In J. C. Trueswell & M. K. Tanenhaus (Eds.), *Approaches to studying world situ-*

ated language use: Bridging the language as product and language as action traditions, Cambridge, MA: MIT Press.

Selten, R., & Warglien, M. (2007). The emergence of simple languages in an experimental coordination game. *PNAS*, 104: 7361-7366.

Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, 133: 140-155.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3): 379-423.

Silvey, C. (2015). *The communicative emergence and cultural evolution of word meanings*. Unpublished PhD Thesis: University of Edinburgh.

Silvey, C., Kirby, S., & Smith, K. (2015). Word meanings evolve to selectively preserve distinctions on salient dimensions. *Cognitive Science*, 39(1): 212-26.

Simonton, D. K. (1975). Galton's problem, autocorrelation, and diffusion coefficients. *Cross-Cultural Research*, 10: 239-248. doi: 10.1177/106939717501000401.

Smith, A. D. M., & Hoffer, S. (2015). The pivotal role of metaphor in the evolution of human language. In J. E. Diaz-Vera (Ed.), *Metaphor and Metonymy Across Time and Cultures: perspectives on the sociohistorical linguistics of figurative language*: 123-140. Berlin: Walter de Gruyter.

Smith, K. (2009). Iterated learning in populations of Bayesian agents. In N.A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, 697-702. Austin, TX: Cognitive Science Society.

Smith, K., & Kirby, S. (2012). Compositionality and linguistic evolution. In W. Hinzen, E. Machery, and M. Werning (Eds.), *Oxford Handbook of Compositionality*. Oxford: Oxford University Press.

Smith, K., Tamariz, M. & Kirby, S. (2013). Linguistic structure is an evolutionary trade-off between simplicity and expressivity. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society*: 1348-1353. Austin, TX: Cognitive Science Society.

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

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science*, 13(1): 139.

Smith, S. M., Glenberg, A., & Bjork, R. A. (1978). Environmental context and human memory. *Memory & Cognition*, 6: 342-353.

Snow, C. E., Barnes, W. S., Chandler, J., Hemphill, L & Goodman, I. F. (1991). *Unfulfilled Expectations: Home and School Influences on Literacy*. Cambridge, MA: Harvard University Press.

Sperber, D. & Wilson, D. (1995/2005). *Relevance: Communication and Cognition*. 2nd Edition. Oxford/Cambridge: Blackwell Publishers.

Spike, M., Stadler, K., Kirby, S., & Smith, K. (2016). Minimal requirements for the Emergence of Learned Signalling. *Cognitive Science*. doi: 10.1111/cogs.12351.

Spirtes, P., Glymour, C. N., & Scheines, R. (2000). *Causation, Prediction, and Search*. Cambridge, MA: MIT press.

Spivey-Knowlton, M., & Tanenhaus, M. (2015). Referential Context and Syntactic Ambiguity Resolution. In C. Clifton, L. Frazier & K. Rayner (Eds.), *Perspectives on Sentence Processing*: Chapter 17. New York, NY: Taylor & Francis.

Sproat, R. (2010). Ancient symbols, computational linguistics, and the reviewing practices of the general science journals. *Computational Linguistics*, 36: 585-594.

Steels, L. (1999). *The Talking Heads Experiment Volume 1: Words and Meanings*. Brussels: Best of Publishing.

Steels, L. (2000). Language as a Complex Adaptive System. In Schoenauer, M., (Ed.), *Proceedings of PPSN VI*. Berlin, Germany: Springer-Verlag.

Steels, L. (2003). Evolving grounded communication for robots. *Trends in Cognitive Sciences*, 7(7): 308-312. doi: 10.1016/S1364-6613(03)00129-3.

Steels, L. (2012). Self-organization and Selection in Cultural Language Evolution. In L. Steels (Ed.), *Experiments in Cultural Language Evolution*, 1-37. Amsterdam: John Benjamins.

Stivers, T., et al. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences of the USA*, 106(26): 10587-10592. doi: 10.1073/pnas.0903616106.

Taleb, N. N. (2012). *Antifragile: things that gain from disorder*. New York, NY: Random House Incorporated.

Tamariz, M. & Kirby, S. (2016). The cultural evolution of language. *Current Opinion in Psychology*, 8: 37-43.

Terkourafi, M. (2009). On de-limiting context. In A. Bergs and G. Dielwald (Eds.), *Contexts and Constructions*: 17-42. Amsterdam: John Benjamins.

Terrell, J. E., Hunt, T. L., & Gosden, C. (1997). Human diversity and the myth of the primitive isolate. *Current Anthropology*, 38: 155-195. doi: 10.1086/204604.

The World Bank. (2013). *World development indicators*. accessed 18/04/2013. <http://data.worldbank.org/indicator/NY.GDP.MKTP.CD/countries>.

Thiesen-White, C., Kirby, S. & Oberlander, J. (2011). Integrating the horizontal and vertical cultural transmission of novel communication systems. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*: 956-961. Austin, TX: Cognitive Science Society.

Thompson, B., Kirby, S., & Smith, K. (2016). Culture shapes the evolution of cognition. *PNAS*, 113: 4530-4535.

Tily, H., & Piantadosi, S. T. (2009). Refer efficiently: Use less informative expressions for more predictable meanings. In *Proceedings of the workshop on the production of referring expressions: Bridging the gap between computational and empirical approaches to reference*.

Todd, M. T., Nystrom, L. E., & Cohen, J. D. (2013). Confounds in multivariate pattern analysis: Theory and rule representation case study. *NeuroImage*, 77: 157-165. doi: 10.1016/j.neuroimage.2013.03.039.

- Todorović, D. (2010). Context effects in visual perception and their explanations. *Review of Psychology*, 17(1): 17-32.
- Tomasello, M. (2008). *Origins of human communication*. Cambridge, MA: MIT Press.
- Tovar, A. (1981). *Relatos y diálogos de los matacos: Seguidos de una gramática de su lengua*. Madrid: Ediciones Cultura Hispánica del Instituto de Cooperación Iberoamericana.
- Traugott, E., & Dasher, R. B. (2002). *Regularity in Semantic Change*. Cambridge: Cambridge University Press.
- Traugott, E., & König, E. (1991). The semantics-pragmatics of grammaticalization revisited. In E. Traugott & B. Heine (Eds.), *Approaches to grammaticalization*: 189-218. Amsterdam: John Benjamins.
- Traugott, E. & Trousdale, G. (2013). *Constructionalization and Constructional Changes*. Oxford: Oxford University Press.
- Trudgill, P. (2004). Linguistic and social typology: The austronesian migrations and phoneme inventories. *Linguistic Typology*, 8: 305-320. doi: 10.1515/lity.2004.8.3.305.
- Trudgill, P. (2011). *Sociolinguistic typology: Social determinants of linguistic complexity*. Oxford: Oxford University Press.
- Tuggy, D. (1993). Ambiguity, polysemy and vagueness. *Cognitive Linguistics*, 4: 273-291.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, 76(2): 105-110.

Ullman, M. (2005). A cognitive neuroscience perspective on second language acquisition: The declarative/procedural model. In C. Sanz, (Ed.), *Mind and Context in Adult Second Language Acquisition: Methods, Theory, and Practice*: 141178. Washington D.C., CA: Georgetown University Press.

United States Senate Committee on the Budget. (2011). *The case for growth: Sessions lists benefits of discretionary cuts*. Online 15/03/2011. <http://www.budgetsenate.gov/republican/public/indexcfm/2011/3/the-case-for-growth-sessions-lists.html>. Accessed 18/04/2013.

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

von Humboldt, W. (1836). *Über die Verschiedenheit des menschlichen Sprachbaues und ihren Einfluss auf die geistige Entwicklung des Menschengeschlechts*. Berlin. Online: <https://books.google.de/books?id=dV4SAAAAIAAJ&printsec=frontcover#v=onepage&q&f=false>

Wade, N. J. & Swanston, M. T. (2001). *Visual Perception: An Introduction*. Philadelphia, PA: Taylor & Francis.

Way, B. M., & Lieberman, M. D. (2010). Is there a genetic contribution to cultural differences? Collectivism, individualism and genetic markers of social sensitivity. *Social Cognitive and Affective Neuroscience*, 5: 203-211. doi: 10.1093/scan/nsq059.

Wedgwood, D. (2007). Shared assumptions: Semantic minimalism and relevance theory. *Journal of Linguistics*, 43: 647-681.

Wichmann, S., Rama, T., & Holman, E. (2011). Phonological diversity, word length, and population sizes across languages: The ASJP evidence. *Linguistic Typology*, 15: 177-197. doi: 10.1515/lity.2011.013.

Wikipedia. (2011). *Siesta*. accessed 01/05/2011. <http://en.wikipedia.org/wiki/Siesta>.

Wikipedia. (2013). *List of sovereign states and dependent territories by population density*. accessed 18/04/2013. http://en.wikipedia.org/wiki/List_of_sovereign_states_and_dependent_territories_by_population_density

Wikipedia. (2013). *List of countries by nobel laureates per capita*. accessed 22/04/2013. http://en.wikipedia.org/wiki/List_of_countries_by_Nobel_laureates_per_capita.

Wikipedia. (2013). *List of countries by gdp (nominal) per capita*. accessed 22/04/2013. http://en.wikipedia.org/wiki/List_of_countries_by_GDP_nominal_per_capita.

Wikipedia. (2013). *List of countries by traffic-related death rate*. accessed 22/04/2013. URL http://en.wikipedia.org/wiki/List_of_countries_by_traffic_related_death_rate.

Wikipedia. (2013). *List of serial killers by number of victims*. accessed 22/04/2013. http://en.wikipedia.org/wiki/List_of_serial_killers_by_number_of_victims.

Wikipedia. (2013). *List of rampage killers*. accessed 22/04/2013. http://en.wikipedia.org/wiki/List_of_rampage_killers.

Windelband, W. (1894/1998). History and natural science. *Theory & Psychology*, 8: 522.

Winter, B., & Ardell, D. (2016). Rethinking Zipf's Frequency-meaning Relationship: Implications for the Evolution of Word Meaning. In S. G. Roberts et al (Eds.) *The Evolution of Language: Proceedings of the 11th International Conference (EVLANG11)*. doi: 10.17617/2.2248195.

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

Wittgenstein, L. (1921). *Tractatus Logico-Philosophicus*. Trans. C. K. Ogden (1922). London: Routledge.

Wittgenstein, L. (1953). *Philosophical Investigations*. G.E.M. Anscombe and R. Rhees (Eds.), Trans. G.E.M. Anscombe. Oxford: Blackwell.

World Health Organization Injuries and Violence Prevention Dept. (2002). *The injury chart book: A graphical overview of the global burden of injuries*. World Health Organization.

Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge, UK: Cambridge University Press.

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

Yanny, B., et al. (2009). Segue: A spectroscopic survey of 240,000 stars with $g=1420$. *The Astronomical Journal*, 137: 4377. doi: 10.1088/0004-6256/137/5/4377.

Yglesias, M. (2013). Is the Reinhart-Rogoff result based on a simple spreadsheet error? *Slate magazine Moneybox blog*. http://www.slate.com/blogs/moneybox/2013/04/16/reinhart_rogoff_coding_error_austerity_policies.html. Accessed 18/04/2013.

Zipf, G. (1949). *Human behavior and the principle of least effort*. New York, NY: Addison-Wesley.

Zhang, Q. (2011). Elasticity of Vague Language. *Intercultural Pragmatics*, 8(4): 571-599.

Appendices

Appendix A: Languages Experiment 1

Initial Languages

The randomly generated initial languages used in the experiment. First column contains the meaning and all subsequent columns are the signals that participants with trained on (organised by chains within the Mixed, Shape-Different, and Shape-Same conditions). To save space, the data is available online at: http://www.replicatedtypo.com/wp-content/uploads/2016/11/exp1_initial.pdf

Mixed Chain 1

Chain 1 in the Mixed condition. The first column contains the meaning, with all subsequent columns being the signal output for participants at each generation and block. To save space, the data is available online at: http://www.replicatedtypo.com/wp-content/uploads/2016/11/exp1_MX1.pdf

Mixed Chain 2

Chain 2 in the Mixed condition. The first column contains the meaning, with all subsequent columns being the signal output for participants at each generation and block. To save space, the data is available online at: http://www.replicatedtypo.com/wp-content/uploads/2016/11/exp1_MX2.pdf

Mixed Chain 3

Chain 3 in the Mixed condition. The first column contains the meaning, with all subsequent columns being the signal output for participants at each generation and block. To save space, the data is available online at: http://www.replicatedtypo.com/wp-content/uploads/2016/11/exp1_MX3.pdf

Mixed Chain 4

Chain 4 in the Mixed condition. The first column contains the meaning, with all subsequent columns being the signal output for participants at each generation and block. To save space, the data is available online at: http://www.replicatedtypo.com/wp-content/uploads/2016/11/exp1_MX4.pdf

Shape-Different Chain 1

Chain 1 in the Shape-Different condition. The first column contains the meaning, with all subsequent columns being the signal output for participants at each generation and block. To save space, the data is available online at: http://www.replicatedtypo.com/wp-content/uploads/2016/11/exp1_SD1.pdf

Shape-Different Chain 2

Chain 2 in the Shape-Different condition. The first column contains the meaning, with all subsequent columns being the signal output for participants at each generation and block. To save space, the data is available online at: http://www.replicatedtypo.com/wp-content/uploads/2016/11/exp1_SD2.pdf

Shape-Different Chain 3

Chain 3 in the Shape-Different condition. The first column contains the meaning, with all subsequent columns being the signal output for participants at each generation and block. To save space, the data is available online at: http://www.replicatedtypo.com/wp-content/uploads/2016/11/exp1_SD3.pdf

Shape-Different Chain 4

Chain 4 in the Shape-Different condition. The first column contains the meaning, with all subsequent columns being the signal output for participants at each generation and block. To save space, the data is available online at: http://www.replicatedtypo.com/wp-content/uploads/2016/11/exp1_SD4.pdf

Shape-Same Chain 1

Chain 1 in the Shape-Same condition. The first column contains the meaning, with all subsequent columns being the signal output for participants at each

generation and block. To save space, the data is available online at: http://www.replicatedtypo.com/wp-content/uploads/2016/11/exp1_SS1.pdf

Shape-Same Chain 2

Chain 2 in the Shape-Same condition. The first column contains the meaning, with all subsequent columns being the signal output for participants at each generation and block. To save space, the data is available online at: http://www.replicatedtypo.com/wp-content/uploads/2016/11/exp1_SS2.pdf

Shape-Same Chain 3

Chain 3 in the Shape-Same condition. The first column contains the meaning, with all subsequent columns being the signal output for participants at each generation and block. To save space, the data is available online at: http://www.replicatedtypo.com/wp-content/uploads/2016/11/exp1_SS3.pdf

Shape-Same Chain 4

Chain 4 in the Shape-Same condition. The first column contains the meaning, with all subsequent columns being the signal output for participants at each generation and block. To save space, the data is available online at: http://www.replicatedtypo.com/wp-content/uploads/2016/11/exp1_SS4.pdf

Appendix B: Languages Experiment 2

Both-Different

Initial languages and participant output for Both-Different condition. Pair refers to the particular participant pair within the condition. The first two columns are the target meaning (shape, colour). Initial is the training language participant pairs were trained on. In the block columns are the signals produced by participant pairs during interaction. For example, Block 1a refers to the first block of interaction and the first set of signals produced for each unique meaning (i.e., a signal for Blob Blue at block 1a preceded the signal for Blob Blue at block 2b). To save space, the data is available online at: http://www.replicatedtypo.com/wp-content/uploads/2016/11/exp2_BD.pdf

Colour-Different

Initial languages and participant output for Both-Different condition. Pair refers to the particular participant pair within the condition. The first two columns are the target meaning (shape, colour). Initial is the training language participant pairs were trained on. In the block columns are the signals produced by participant pairs during interaction. For example, Block 1a refers to the first block of interaction and the first set of signals produced for each unique meaning (i.e., a signal for Blob Blue at block 1a preceded the signal for Blob Blue at block 2b). To save space, the data is available online at: http://www.replicatedtypo.com/wp-content/uploads/2016/11/exp2_CD.pdf

Shape-Different

Initial languages and participant output for Shape-Different condition. Pair refers to the particular participant pair within the condition. The first two columns are the target meaning (shape, colour). Initial is the training language participant pairs were trained on. In the block columns are the signals produced by participant pairs during interaction. For example, Block 1a refers to the first block of interaction and the first set of signals produced for each unique meaning (i.e., a signal for Blob Blue at block 1a preceded the signal for Blob Blue at block 2b). To save space, the data is available online at: http://www.replicatedtypo.com/wp-content/uploads/2016/11/exp2_SD.pdf

Appendix C: Languages Experiment 3

Both-Different

Initial languages and participant output for Both-Different condition. Pair refers to the particular participant pair within the condition. The first two columns are the target meaning (shape, colour). Initial is the training language participant pairs were trained on. In the block columns are the signals produced by participant pairs during interaction. For example, Block 1a refers to the first block of interaction and the first set of signals produced for each unique meaning (i.e., a signal for Blob Blue at block 1a preceded the signal for Blob Blue at block 2b). To save space, the data is available online at: http://www.replicatedtypo.com/wp-content/uploads/2016/11/exp3_BD.pdf

Colour-Different

Initial languages and participant output for Colour-Different condition. Pair refers to the particular participant pair within the condition. The first two columns are the target meaning (shape, colour). Initial is the training language participant pairs were trained on. In the block columns are the signals produced by participant pairs during interaction. For example, Block 1a refers to the first block of interaction and the first set of signals produced for each unique meaning (i.e., a signal for Blob Blue at block 1a preceded the signal for Blob Blue at block 2b). To save space, the data is available online at: http://www.replicatedtypo.com/wp-content/uploads/2016/11/exp3_CD.pdf

Shape-Different

Initial languages and participant output for Shape-Different condition. Pair refers to the particular participant pair within the condition. The first two columns are the target meaning (shape, colour). Initial is the training language participant pairs were trained on. In the block columns are the signals produced by participant pairs during interaction. For example, Block 1a refers to the first block of interaction and the first set of signals produced for each unique meaning (i.e., a signal for Blob Blue at block 1a preceded the signal for Blob Blue at block 2b). To save space, the data is available online at: http://www.replicatedtypo.com/wp-content/uploads/2016/11/exp3_SD.pdf

Appendix D: Languages Experiment 4

Initial Languages

The initial languages used in Experiment 4. The first column is the randomly generated language and the second column is the meaning. Each initial language consists of 4 signals paired with 4 meanings. To save space, the data is available online at: http://www.replicatedtypo.com/wp-content/uploads/2016/11/exp4_initial.pdf

Shape-Different + Shared Context Output

The output for the Shape-Different + Shared Context condition. Pair refers to the particular pair involved in interaction (which, in this case, consists of fixed roles for speakers and hearers). The first column is the target meaning a speaker needed

to convey. Distractor 1, 2 and 3 is the referential context within which the target was situated: this is what the hearer saw when they needed to select an image. Description is the signal produced by the speaker. Data is sorted according to target. To save space, the data is available online at: <http://www.replicatedtypo.com/wp-content/uploads/2016/11/SD-Shared-Context.pdf>

Shape-Different + Unshared Context Output

The output for the Shape-Different + Unshared Context condition. Pair refers to the particular participant pair involved in interaction (which, in this case, consists of fixed roles for speakers and hearers). Each separate table within a pair refers to a block of interaction. The first column is the target meaning a speaker needed to convey. Distractor 1, 2 and 3 is the referential context within which the target was situated: this is what the hearer saw when they needed to select an image. Description refers to the signal produced by a speaker. Data is sorted alphabetically according to the target. To save space, the data is available online at: <http://www.replicatedtypo.com/wp-content/uploads/2016/11/SD-Unshared-Context.pdf>

Mixed + Shared Context Output

The output for the Mixed + Shared Context condition. Pair refers to the particular participant pair involved in interaction (which, in this case, consists of fixed roles for speakers and hearers). Each separate table within a pair refers to a block of interaction. The first column is the target meaning a speaker needed to convey. Distractor 1, 2 and 3 is the referential context within which the target was situated: this is what the hearer saw when they needed to select an image. Description refers to the signal produced by a speaker. Data is sorted alphabetically according to the target. To save space, the data is available online at: <http://www.replicatedtypo.com/wp-content/uploads/2016/11/MX-Shared-Context.pdf>

Mixed + Unshared Context Output

The output for the Mixed + Unshared Context condition. Pair refers to the particular participant pair involved in interaction (which, in this case, consists of fixed roles for speakers and hearers). Each separate table within a pair refers to a block of interaction. The first column is the target meaning a speaker needed to convey. Distractor 1, 2 and 3 is the referential context within which

the target was situated: this is what the hearer saw when they needed to select an image. Description refers to the signal produced by a speaker. Data is sorted alphabetically according to the target. To save space, the data is available online at: <http://www.replicatedtypo.com/wp-content/uploads/2016/11/MX-Unshared-Context.pdf>