



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.

The communicative emergence and cultural evolution of word meanings

Catriona Silvey



Doctor of Philosophy
School of Philosophy, Psychology and Language Sciences
University of Edinburgh
2014

Abstract

The question of how language evolved has received an increasing amount of attention in recent years. Compared to seemingly more complex phenomena such as syntax, word meanings are usually seen as relatively easy to explain. Mainstream accounts in psycholinguistics and evolutionary linguistics assume that word meanings correspond to stable concepts which are prior to language and derive straightforwardly from human perception of structure in the world. Taking a cognitive linguistic approach based on psycholinguistic evidence, I argue instead that word meanings are conventions, grounded, learned and used in the context of communication. The meaning of a word is the sum of its contexts of use, with particular features of these contexts made more or less salient by mechanisms of attentional learning and communicative inference. Evolutionarily, word meanings arise as an emergent product of humans' adapted tendency to infer each other's intentions using contextual cues. They are then shaped over cultural evolution by the need to be learnable and useful for communication. This thesis presents a series of experiments that test the effect of these pressures on the origins and development of word meanings. Experiment 1 investigates the origins of strong tendencies for words to specify features on particular dimensions (such as the shape bias). The results show that these tendencies arise via attentional learning effects amplified by iterated learning. Dimensions which are less salient in contexts of learning and use drop out of word meanings as they are passed down a chain of learners. Experiments 2, 3 and 4 investigate the structure of word meanings produced during either paired communication games or individual labelling of images by similarity. While communication alone leads to word meanings that are unstructured and poorly aligned within pairs, communication plus iterated learning leads to word meanings that increase in structure and alignment over generations. Finally, Experiment 5 investigates the interaction of event structure and developing conventions in shaping word meanings. The structure of events in an artificial world is shown to influence lexicalisation patterns in the languages conventionalised by communicating pairs. Event features that are less predictable across communicative contexts tend to be more strongly associated with the conventions in the language. Overall, the experiments show that rather than straightforwardly reflecting pre-linguistic conceptualisation, word meanings are also dynamically shaped by learning and communication. In addition, these processes are constrained by the conventions that already exist within a language. This illuminates the mixture of convergence and diversity we see in word meanings in natural languages, and gives insight into their evolutionary origins.

Lay Summary

Human language is unique in the natural world. By using words – a set of learned sounds that have meanings – we can communicate our intentions and infer the intentions of others. Where do word meanings come from? Mainstream work in evolutionary linguistics argues that word meanings existed before language: words label concepts that correspond to categories of objects and actions in the world. I argue instead that word meanings are shaped by pressures specific to language: cultural transmission (since children learn word meanings from adults) and communication (since we use words as tools for communicating and inferring intentions). This implies that word meanings are shaped by what humans are capable of learning from limited input, and inferring from communicative contexts.

I present the results of artificial language experiments that test the effects of cultural transmission and communication on the structure of word meanings. The results show that cultural transmission and communication have important and differing effects on word meaning structure. Cultural transmission makes word meanings more learnable by reducing the number of meaning distinctions and directing these distinctions to important dimensions. For example, if shape is more important than colour, words in a language will keep shape distinctions and lose colour distinctions over generations of transmission. Communication works on top of this learned system, shaping word meanings to be efficient clues for inferring a speaker's intention given the structure of the world, and leading word meanings to be shared between communicators. Overall, the results support a picture where word meanings, rather than simply being concepts that exist prior to language, are an evolved compromise satisfying the need for language to be both learnable and useful for communication.

Acknowledgements

Thanks and no thanks to Robin Hobb for releasing the first volume of a new trilogy in the last month of my writeup, thus ensuring substantial portions of this thesis were written through a haze of tears, and to the BBC tent at the Edinburgh Festival Fringe, for ensuring the rest was written to a thumping bass soundtrack.

Thanks to the Carnegie Trust and the AHRC for funding me – I hope I’ve been a good investment.

My supervisors have been the best possible balance of encouraging and critical. Thanks to them I went from feeling like a total impostor to feeling like an impostor who had successfully persuaded people that I knew what I was doing. Thanks to Kenny Smith, my awesome lead supervisor, for generally keeping me sane, for being an experimental design and paper-writing genius, and for putting up with frequent random interruptions involving panicked questions to which he always responded calmly and insightfully. Simon Kirby has been a fantastic second supervisor, always on hand with alternative interpretations of results and big-picture questions that were just the right mix of reassuring and unsettling. Thanks also to Nik Gisborne for helpful discussions and for keeping my semantics in line.

I still can’t believe the LEC let in an English Lit graduate who started wondering where all these words came from. Thanks to all the smart and hilarious PhD students, postdocs and staff who make it such a great place to work.

The residents of office 1.15 have been very tolerant of me regularly throwing off my headphones and shouting things like ‘What is learning??’ Particular thanks to Yasamin Motamedi, Mark Atkinson, Carmen Saldaña, James Winters, Thijs Lubbers, Matt Spike, Kevin Stadler, Vanessa Ferdinand, Bill Thompson, Alan Nielsen, Ruth Friskney, George Starling, Steph DeMarco, Soundess Azzabou-Kacem, and Rea Colleran. Special thanks also to Hannah Little for always being around on Google chat to discuss burning questions such as whether bees have inference.

Huge thanks go to my mum and dad, without whose financial support I could not have done the MSc, and without whose support of every other kind I definitely could not have finished the PhD. Sorry for throwing away a lucrative career in non-profit science publishing for possibly the only less lucrative option available.

And lastly thanks to Christos, for being my Greek informant and my coding tutor, for his endless patience with my meltdowns, his constant willingness to have arguments about semantics, and his never-failing ability to make me laugh. Σ’ευχαριστώ τόσο πολύ. Σ’αγαπώ.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(Catriona Silvey)

Contents

1	Introduction	1
1.1	Aim of the thesis	3
1.2	What are the meaning units of language?	3
1.2.1	Words	4
1.2.2	Below the word level	5
1.2.3	Above the word level	5
1.3	Investigating meaning using artificial language experiments	7
1.4	Thesis road map	8
2	The description of word meanings	11
2.1	Introduction	11
2.2	Word meanings as stable pre-linguistic concepts	13
2.2.1	Problem 1: Cross-linguistic variation	14
2.2.2	Problem 2: Contextual variation	20
2.3	Word meanings as emergent from utterance comprehension	26
2.3.1	Problem 1: Systematicity	30
2.3.2	Problem 2: Compositionality	31
2.4	Conclusion	32
3	The evolution of word meanings	33
3.1	Introduction	33
3.2	Theories of word meaning origin	34
3.2.1	The protolanguage debate	36
3.2.2	Propositional meaning	37
3.2.3	Emergent meaning	38
3.3	Models of word meaning	40
3.3.1	Models with built-in meanings	40

3.3.2	Models with constructed meanings	42
3.4	Experiments on word meaning	46
3.4.1	Perceptual meanings	46
3.4.2	Task-defined meanings	48
3.5	Pressures and mechanisms	50
3.5.1	World structure	51
3.5.2	Learning	52
3.5.3	Communication	54
3.6	Hypotheses	57
3.7	Conclusion	58
4	The origins of underspecification	61
4.1	Introduction	61
4.2	Motivation	62
4.2.1	Attentional learning and the shape bias	62
4.2.2	Modelling the cultural evolution of underspecification: iterated learning	64
4.3	Method	66
4.3.1	Participants	66
4.3.2	Stimuli: images and input language	66
4.3.3	Procedure	66
4.3.4	Dependent variables	70
4.4	Results	73
4.4.1	Transmission error	73
4.4.2	Underspecification	74
4.4.3	Language structure	82
4.5	Discussion	83
4.5.1	Underspecification on backgrounded dimensions	83
4.5.2	Partial compositionality	84
4.5.3	Language structure	87
4.5.4	Problems and future directions	88
4.6	Conclusion	90
5	Learning, communication, and the structure of word meanings	93
5.1	Introduction	93
5.2	General Method	95

5.2.1	Stimuli	95
5.2.2	Outcomes	96
5.2.3	Experiment 2	103
5.3	Experiment 3	108
5.3.1	Method	108
5.3.2	Experiment 3 Results	112
5.3.3	Experiment 3 Discussion	119
5.4	Experiment 4	122
5.4.1	Method	122
5.4.2	Experiment 4 Results	125
5.4.3	Experiment 4 Discussion	138
5.5	Exploratory results	140
5.5.1	Factors affecting communicative success	140
5.5.2	Consensus on category boundaries	144
5.6	Design issues and future directions	146
5.7	Conclusion	149
6	The interaction of word meanings with event structure	153
6.1	Introduction	153
6.2	Background	155
6.3	Methods	156
6.3.1	Participants	157
6.3.2	Stimuli	158
6.3.3	Training language	161
6.3.4	Procedure	162
6.3.5	Dependent variables	166
6.4	Hypotheses	168
6.4.1	Predictable versus unpredictable	168
6.4.2	Objects versus actions	170
6.5	Results	172
6.5.1	Event structure test	172
6.5.2	Communicative success	174
6.5.3	Event feature encoding	175
6.6	Discussion	177
6.6.1	Event test	177

6.6.2	Communicative success	178
6.6.3	Event feature encoding	178
6.6.4	Strategies	180
6.6.5	Problems and future directions	184
6.7	Conclusion	186
7	Conclusion	187
A	Published Papers	195
	Bibliography	217

Introduction

Defining the meaning of a word is an enterprise of almost inconceivable complexity.

Elbourne (2011, p. 1)

Vizzini: Inconceivable!

Inigo: You keep using that word. I do not think it means what you think it means.

The Princess Bride (1987)

Language is an extremely versatile tool for communication. Using language, humans are able to share knowledge about events, beliefs, attitudes, and plans by uttering structured sequences of sounds. This ability is unparalleled in the natural world. While other animals possess simple communication systems – for example, the alarm calls of vervet monkeys, which distinguish between different types of predator (Seyfarth et al., 1980), or the waggle dances of bees, which communicate direction and distance of sources of nectar from the hive (von Frisch, 1967) – these are closed signal repertoires whose use is bound to limited contexts. Language, by contrast, is productive, in the sense that we are capable of producing and understanding utterances we have never heard before. For example, if I tell you that when my granny first met her father-in-law she brought him a bottle of whisky and he poured it down the sink, you can build up

a fairly accurate model of this event and even make inferences about the respective beliefs of its participants, regardless of how atypical it might be of your experiences of whisky, sinks or grandmothers, and regardless of whether you have heard those words in exactly that combination before.

Linguists generally agree that we accomplish this trick by means of two main kinds of knowledge. The first is lexical knowledge, whereby we know the meanings of the individual words and morphemes that make up an utterance; the second is grammatical knowledge, whereby we know the rules by which these meanings should be combined (Pinker, 1999). The productivity of language is therefore argued to be dependent on its compositionality: the meaning of an utterance is a function of the meanings of its parts (determined by lexical knowledge) and the way those parts are combined (determined by grammatical knowledge).

In recent years, an increasing amount of attention has been paid to the question of how language evolved (for recent conference proceedings see Smith et al., 2010a; Scott-Phillips et al., 2012; Cartmill et al., 2014). Much of this work has focused on the origins of the various levels of structure in which linguistic units are combined, in particular syntax and morphology – grammatical knowledge, by the definition above.¹ However, the origins of lexical knowledge have generally received less attention. There are two main reasons for this. On the one hand, in principle, lexical knowledge has been seen as relatively easy to explain: word meanings are modelled as corresponding to parts of the world or their mental representations, and researchers often choose to abstract away from the precise workings of this relationship in order to focus on other questions. On the other hand, in practice, unpacking the relationship between language, the world, and cognition and investigating it experimentally is challenging, given the difficulties of characterising word meanings even in natural languages (see above epigraph from Elbourne, 2011). However, despite these difficulties, meaning is what distinguishes language from even the more structured forms of animal communication such as bird and whale song, and enables its role in supporting rich cumulative culture. Charles Hockett, in his article ‘The Origins of Speech’, characterises ‘semanticity’, or ‘relatively fixed associations between elements in messages (e.g., words) and recurrent features or situations of the world around us’, as one of the design features of language (Hockett, 1960, p. 6). This semanticity is also key to the design feature of productivity: if ‘one can coin new utterances by putting together pieces familiar from

¹There has also been a substantial amount of work on phonology, both in isolation and in relation to morphosyntax: see de Boer et al. (2012) for a recent special issue on this topic.

old utterances', and these new utterances can be 'understood by other speakers of the language' (Hockett, 1960, p. 6), this relies on the familiar utterance pieces having associated meanings. Understanding what the meaning units of language are, and how they arise, develop and are cognitively represented, is therefore crucial to understanding the origins of language.

1.1 Aim of the thesis

In this thesis I aim to investigate the origins and evolution of word meanings, defined here as the lexical knowledge which, in conjunction with grammatical knowledge, allows us to produce and comprehend novel utterances. Using artificial language experiments, I explore how word meanings arise and change as participants learn and coordinate on shared communicative conventions. In particular, I will investigate how habitual features of learning and production context cause particular meaning distinctions to be maintained at the expense of others; how learning and communication interact over cultural transmission to increase the structure of word meanings and their alignment within interacting pairs; and how patterns of event structure influence the development of word meanings, driving the lexicalisation of event features that are less predictable from linguistic and non-linguistic context. The overall aim is to model the pressures that might have shaped the semantic systems we see in the world's languages, and to investigate how these pressures interact with the developing structure of a given system, shaping new meanings to fit in with existing conventions.

The rest of this introduction will ask what the meaning units of language are, and motivate using artificial language experiments to investigate their origins and development. Finally, I will outline a road map of the rest of the thesis.

1.2 What are the meaning units of language?

In the standard account of language production and comprehension sketched above, there is a clear distinction between knowledge of the meaning-bearing units of a language and knowledge of the rules by which these units are combined. However, in practice, the distinction is somewhat blurry. Intuitively, meaning-bearing units should be identifiable by virtue of 'carrying a constant meaning from context to context' (Cruse, 2011, p. 84); however, it is surprisingly difficult to isolate a linguistic level at which this consistently applies.

This section will first make the assumption that the meaning-bearing units of language are words, then look for evidence of meaning units below and above the word level.

1.2.1 Words

Words, words. They're all we have to go on. (Tom Stoppard, *Rosencrantz and Guildenstern Are Dead*)

The obvious candidate for the meaning-bearing unit of language is the word. Intuitively, we think of language as being made up of words: not only are they visibly separated in writing, but in spoken language they are 'structurally inviolable wholes', moveable relatively flexibly within an utterance, but typically not split or interrupted (Cruse, 2011, p. 75).

However, despite these intuitions, there is reason to doubt that words pick out a meaningful level of linguistic structure. Haspelmath (2011) surveys ten different proposed criteria for determining what a word is, and finds them all lacking: 'there is no definition of 'word' that can be applied to any language and that would yield consistent results' (Haspelmath, 2011, p. 28). Haspelmath argues that the concept of a word only emerged relatively recently, as an artefact of alphabetic writing systems: 'In European languages...we can see that the current words for 'word' (e.g. English *word*, French *mot*, Russian *slovo*, etc.) originally started out with much more general meanings ('act of speaking') and seem to have acquired the narrower sense of Greek *léxis* only through formal schooling, in particular writing and grammar teaching' (Haspelmath, 2011, p. 33).

Looking specifically at the proposed semantic criteria for wordhood, some authors have defined a word as the smallest meaningful unit of language. For example, Sapir (1921) characterises words as 'the smallest, completely satisfying bits of isolated 'meaning' into which the sentence resolves itself' (p. 34). This includes words usually characterised as semantically complex, e.g., 'unthinkable'. Sapir argues that even these complex words 'cannot be cut into without a disturbance of meaning, one or the other or both of the severed parts remaining as a helpless waif on our hands'. While 'think' can stand on its own, 'un-' and '-able' cannot, leaving 'unthinkable' as a word by Sapir's criteria. However, this definition fails to account for the productivity of 'un-' and '-able': they can attach to other verb stems in a parallel way to express a parallel meaning, e.g., 'unclimbable', 'undrinkable'. This strongly suggests that these

sub-word elements carry independent meaning. For this reason, most researchers agree that to find the meaning units of language, we have to look below the word level.

1.2.2 Below the word level

Perhaps there is no way out: there are just so many goddamned words, and so many parts to them. (Ray Jackendoff, *Foundations of Language*)

That meaning-bearing units exist below the word level is uncontroversial in linguistics. These units are morphemes, defined as ‘the meaningful units of word building in a language’ (Akmajian et al., 2001, p. 17). This category includes some words, composed of single morphemes (e.g., ‘tree’). However, other words are composed of multiple morphemes (e.g., ‘trees’). Some morphemes can occur in isolation (free morphemes) and some can only appear in conjunction with a restricted set of others (bound morphemes, such as ‘-s’ in the previous example). The constraints on the appearance of bound morphemes relate to their semantic properties: for example, the ‘-ed’ past tense marker co-occurs only with verbs because the time at which an event took place is a saliently specifiable aspect of verb meaning (Cruse, 2011, p. 268).

Whether a word consists of one or several morphemes, however, is not always clear-cut. Regular affixes such as past tense and plural markers constitute one end of the continuum. In the middle are noun-noun compounds, where the contribution of the meanings of the parts to the overall meaning is not always systematic: ‘there is no obvious way of predicting that, for instance, a *tablecloth* is used to cover a table, but a *dishcloth* is used to wipe dishes’ (Cruse, 2011, p. 89). At the other end of the continuum are cases of frozen compositionality, where multiple morphemes can be discerned in a word (and were presumably involved in its derivation), but speakers are not typically aware of the word as multimorphemic. This suggests that the word’s meaning has become independent of the meanings of its original parts. This can be the case even if the units are free morphemes in other contexts: for example, ‘understand’, ‘cupboard’, ‘suitcase’ (Wray & Grace, 2007).

1.2.3 Above the word level

And yet, oh, and yet, we, all of us, spend all our days saying to each other the same things time after weary time: “I love you”, “Don’t go in there”, “Get out”, “You have no right to say that”, “Stop it”, “Why should I”, “That hurt”, “Help”, “Marjorie is dead”. (*A Bit of Fry & Laurie*)

Just as observable meaning units exist below the level of the word, they also exist above. The most often cited are non-transparent idioms, such as ‘kick the bucket’, which have been argued to be treated as single words in processing terms (Estill & Kemper, 1982; Gibbs, 1985). However, this phenomenon is not limited to a few exceptions. Again, as with meaning below the word level, these cases may actually be on a continuum. Some idioms are more semantically transparent than others, in that their parts correspond more clearly to the parts of their meaning (e.g., in ‘pop the question’, ‘pop’ corresponds to ‘suddenly make’ and ‘the question’ to ‘a marriage proposal’, Gibbs et al., 1989). Further along the continuum are collocations and clichés, where the meanings of the units still contribute to the meaning of the whole, but the phrase may have picked up additional meaning of its own, in a phrasal parallel to the noun-noun compounds discussed above. Cruse argues that ‘fish and chips’ falls into this borderline zone: while ‘fish’ and ‘chips’ contribute to the meaning of the phrase, the phrase carries additional specifications, for example that the fish must be battered (Cruse, 2011, p. 92). Frozen oven chips and raw sardines, for example, would count as ‘chips and fish’, but arguably not as ‘fish and chips’.

These phrases are similar to the noun-noun compounds discussed above. In both cases, the meaning of the word or phrase appears simultaneously to be composed of meaning units, and to be a meaning unit itself. In the words of Ruth Millikan, the meaning here is ‘simultaneously derived from two sources, (1) derived compositionally and (2) resulting directly from its holistic reproduction to serve that same function’ (Millikan, 2001, p. 407). Whether the word or phrase is processed as a single meaning-bearing unit ‘may be a matter of degree, or may vary from occasion to occasion’ (p. 407). Elaborating on this point, Alison Wray notes that different users of a language may vary in whether they understand a given phrase holistically, or as built compositionally out of parts. For example, British English speakers using the phrase ‘take a rain check’ may vary in whether they take it as an unanalysed idiom or think (wrongly, in terms of the US English origin of the phrase) that it actually involves checking for rain (Wray, 2002, p. 60). This indeterminacy can lead to gradual diachronic change, where a phrase that was initially unambiguously compositional now seems unambiguously holistic: ‘The meaning of a whole easily separates from the compositional meaning that would be derived from its parts, and may then evolve independently, as in the slippage from...“God be with you” to “Goodbye”’ (Millikan, 2001, p.406). This ‘slippage’ is only possible because during the intermediate phase, the meaning of the phrase is simultaneously compositional and holistic: i.e. it is both composed out of meaning

units and a meaning unit itself.

Taking all of this together, the meaning units of language may not be objectively definable. Chapters 2 and 3 build on this to argue that meaning units in a language, rather than existing a priori, emerge from the processes of learning and communication as language is transmitted and used. In this model, the meaning of a word or an expression is equated to the range of contexts, linguistic and non-linguistic, in which it can be reused. In the remainder of the thesis, however, ‘word’ and ‘word meaning’ will be used as a conventional way of referring to the signal units and meaning units being investigated.

The next section will outline how artificial language learning experiments can be a useful tool for investigating the effects of learning and communication on word meanings, and how these methods will be applied to the questions investigated in this thesis.

1.3 Investigating meaning using artificial language experiments

Word meaning is something of a paradox. Adults within a speech community agree on the meanings of words to the extent that they can use them to communicate successfully with each other most of the time. Children learn the meanings of words very quickly, sometimes after only one exposure (Bloom, 2000). However, describing and formally investigating word meanings in natural languages is notoriously hard. The problem is that many of the relevant facts are accessible only through intuitions that may be unreliable and subject to bias: ‘Semantics is difficult, because unlike phonetic substance, semantic substance cannot be measured or observed objectively’ (Haspelmath, 2000, p. 26). Native speaker intuitions are therefore often supplemented by experimental methods such as semantic acceptability judgements, priming, lexical decision tasks, eyetracking or ERP studies. However, these methods are usually employed for descriptive work on how meanings are represented in an already-developed language. The questions to be explored in this thesis require investigation of how pressures from language learning and use affect the structure of word meanings, ideally without interference from an established language.

Artificial language learning experiments have a long history in psychology and linguistics. Designing artificial languages allows the experimenter to abstract away the features of interest for investigation, without interference from the full complexity of

natural language. This approach has led to insights into a wide range of linguistic phenomena, from the generalisation of novel nouns (Landau et al., 1988) to grammar learning (Reber, 1967) to the segmentation of speech into words (Saffran et al., 1996); for a review, see Folia et al. (2010). For the field of language evolution, artificial language experiments have been a particularly fruitful methodology. Here, the choice of particular input languages, stimuli and tasks has allowed the investigation of phenomena from compositionality (Kirby et al., 2008) to regularisation (Smith & Wonnacott, 2010) to combinatorial phonology (Verhoef et al., 2014). Meanwhile, categorisation experiments using artificial stimuli are a long-established method in the psychological concept learning literature, revealing important insights about the learning and cognitive representation of concepts (for a review, see Murphy, 2002). For example, Rosch & Mervis (1975) ran complementary experiments using natural and artificial categories to explore aspects of prototype structure. The work presented in this thesis brings together these two threads of research, adapting methodologies from experimental research in language evolution that have so far mainly been employed to investigate signal structure to investigate the structure of word meanings.

The obvious advantage of using artificial language experiments to investigate the origins and development of word meanings is that the data are clean and easily quantifiable. The controlled experimental contexts abstract away from much of the indeterminacy of linguistic meaning, allowing confidence about the extent to which participants' behaviour (i.e., the contexts in which they choose to reuse labels) reflects their intuitions about meaning in these tasks. On the other hand, the data that come out of these experiments are impoverished relative to the richness of real linguistic data: the inferences or generalisations participants can make are constrained by the experimental tasks. The experimental approach is therefore a tradeoff, gaining greater control and more quantifiable results at the risk of missing some of the complexity of meaning in natural language. Chapter 2 will further motivate, on theoretical grounds, the decision to treat participants' patterns of word reuse as meanings.

1.4 Thesis road map

The structure of the thesis is as follows.

Chapter 2 will approach the question of what word meanings are and how they are cognitively represented. Starting from the position that word meanings correspond to concepts that exist prior to and independently of language, the chapter will review

evidence from linguistics and psychology that challenges this view. I conclude by adopting a view where word meanings are conventions, grounded, learned and used in the context of communication. The meaning of a word is the sum of its contexts of use, with generalisations to new uses made on the basis of features of the exemplars that are more salient in linguistic and non-linguistic context.

Chapter 3 expands on this view from an evolutionary perspective to ask what pressures we might expect to affect the structure of word meanings, and how we can model these pressures experimentally. Reviewing previous approaches to meaning in the language evolution literature, I argue for two main influences on word meaning structure: human learning biases and communicative inference mechanisms. I then outline how the experiments presented in the thesis will investigate the effects of these pressures.

Chapter 4 presents Experiment 1, which investigates the origin of underspecification: the tendency for words to apply to referents that share features on some dimensions (e.g., shape) but differ on others (e.g., colour). These strong tendencies to specify on particular dimensions emerge gradually from a language which originally specifies across all dimensions equally, as a result of particular dimensions being more or less habitually salient in learning and production contexts. This result shows that a lexicon that supports the learning of higher-order generalisations – e.g., the fact that count nouns tend to specify shape – can itself arise as the product of individuals' lower-level generalisations amplified by cultural transmission.

Experiments 2, 3 and 4, presented in Chapter 5, turn to a stimulus space which lacks discrete feature-based structure. Participants create word meanings by applying the same or different labels to images drawn from this space. Experiment 2 establishes a baseline similarity-based categorisation of the images by collecting participants' pairwise similarity ratings. Experiment 3 investigates whether communication incentivises different word meaning structures from individual categorisation. Surprisingly, word meanings produced by communicators are structured less optimally for communication than those produced by non-communicating individuals. Furthermore, communicating pairs show lower levels of word meaning alignment than matched pairs of non-communicating individuals. This sub-optimality appears to stem from the difficulty of coordinating on word meanings from scratch in a novel space, leading participants to agree on conventions for salient exemplar images, rather than developing optimally structured and aligned categories. To investigate this further, Experiment 4 adds in cultural transmission. Each communicating pair first learns a system of word meanings. For the first pair in each chain, this is a system where each image has a unique label;

for subsequent pairs, it is the system produced by the previous pairs in the chain during communication. Each pair then communicates in turn using the system they have learned, and passes their word meanings on to the next pair. Over generations, this combination of learning and communication leads to word meanings that are better structured for communication and more aligned within pairs. This suggests that both learning and communication are necessary for word meanings to be structured and aligned within a speech community. Learning works to reduce the number of meanings to be learned and rationalise their structure, and communication works to further optimise this structure for communicative success and bring pairs' word meanings into alignment.

Chapter 6 presents Experiment 5, which investigates how communicating pairs generalise a small learned lexicon to describe complex events in an artificial world. The experiment investigates whether constraints on possible event structures affect the event features that are lexicalised in the emerging languages. These constraints do not have a strong effect during communication, where conventionalisation is erratic and noisy. However, a post-test where participants generalise their communicative system shows effects of event structure constraints on participants' word meanings. Event features that are more predictable from linguistic and non-linguistic context are lexicalised less strongly. The results illuminate the interplay between communicators' common ground knowledge of event structure and the emerging conventions of a language, showing that the word meanings which emerge are dependent on both.

The final chapter summarises the work presented in the thesis and suggests future research directions.

The description of word meanings: Language, the mind, and the world

2.1 Introduction

What are word meanings¹, and how are they learned, used, and represented? This chapter will begin by taking an intuitive and widely held view: word meanings correspond to stable concepts which are prior to language and derive straightforwardly from human perception of structure in the world. I raise two main classes of objection to this view. Firstly, the meanings of words exhibit wide variation across languages. Secondly, the meaning of a word varies within a language as a function of linguistic and non-linguistic context. I will consider a number of proposed solutions to these problems and conclude that, rather than being discrete chunks of conceptual structure that exist prior to language, word meanings are conventions that emerge as a by-product of the comprehension of utterances. In this model, a word meaning is represented by exemplars of its contexts of use, with generalisations to new uses made flexibly on the basis of features made salient by linguistic and non-linguistic context. In this way, the uses of a given word are continually reshaped over episodes of learning and communication. The aim of this chapter is twofold: 1) to argue that word meanings do not

¹As discussed in Chapter 1, meaning units in language appear to exist both above and below the word level, and may not be objectively defineable. However, 'word meaning' is used throughout this thesis as a term of convenience.

simply reflect world structure as filtered through human perception, but are shaped by language learning and communication, and 2) to justify taking patterns of word reuse as diagnostic of meaning in the experiments presented later in the thesis. Chapter 3 will then outline the implications of this account for how word meanings could have evolved, and expand on how the experiments will test the effects of language learning and communication on the structure of word meanings.

When we use language, we are usually communicating about ‘things, happenings, and states of affairs in the world’ (Cruse, 2011, p. 46). As such, a reasonable first assumption is that words refer to real-world objects, actions, and properties. If we are in a room with a horse, a cat, and a Golden Retriever, and I say ‘dog!’, you can reasonably infer that I intend to direct your attention to the Golden Retriever. However, this correspondence between words and phenomena in the world is generally not one-to-one.² I can use ‘dog’ to talk about the retriever, but I can also talk about Lassie the fictional collie, or Daisy, my friend’s Schnauzer, who died of extreme stupidity some years ago. Note that some of these dogs no longer exist in the world (Daisy), or indeed never existed at all (Lassie). The problem with characterising word meaning referentially, as a correspondence between a word and a set of real objects, is that if meaning is defined in this extensional way, the meaning of the word ‘dog’ changed when Daisy died. This, while touching, is intuitively problematic.

Formal semantics solves this problem by positing ‘possible worlds’ to define the intension, rather than the extension, of a word. For example, the intension of ‘dog’ is the set of all dogs in all possible worlds, i.e., all the dogs there ever were, are, or potentially could be (Kearns, 2000). However, this account does not get us closer to determining how word meanings are cognitively represented. As Gregory Murphy points out, ‘people do not know or have access to these sets of objects’ (Murphy, 2002, p. 387). Cognitively oriented models of semantics tend to argue instead that words meanings are concepts: abstract mental representations of categories of phenomena in the world (Murphy, 2002). Thus, the connection between words and the world is mediated through mental representations (Ogden & Richards, 1923). The question of word meaning then becomes: what are the concepts associated with words, and how are they learned and used via experience with the world and with language?

²An exception is proper names, which serve to uniquely pick out a particular referent in the world. The semantics of proper names has been a matter of much debate: see Cumming (2013) for discussion.

2.2 Word meanings as stable pre-linguistic concepts

One influential view, associated with Steven Pinker among others, is that word meanings correspond to concepts which the child acquires prior to acquiring language: ‘there really are things and kinds of things and actions out there in the world, and our mind is designed to find them and to label them with words’ (Pinker, 1994, p. 154). Versions of this view differ in the extent to which these pre-linguistic concepts are characterised as innate, versus learned from early experience. The extreme innatist version comes from Jerry Fodor, who argues that the meanings of words, up to and including DOORKNOB, are ‘an inventory of primitive concepts’ (Fodor, 1998, p. 27) which, rather than being learned, are ‘triggered’ by experience. However, most theorists stop short of this position. Ray Jackendoff argues that lexical concepts are composed of semantic primitives, with learning consisting of ‘putting meanings together from smaller parts’ (Jackendoff, 2002, p. 334); Pinker holds a similar view. Paul Bloom, meanwhile, holds that concepts are constructed in support of explanatory theories, with their features selected by means of a possibly innate essentialism, or belief that perceptual properties are explained by deep causal properties (Bloom, 2000).

Whether the concepts concerned are atomic, composed out of innate primitives, or constructed on the basis of essentialist theories, the result is that a word comes to be associated with a concept that would exist regardless of this label: ‘With or without language, the mind has to have a way to unify multimodal representations and store them as units...The structures that make this a “lexical item” rather than just a “concept” simply represent an additional modality into which this concept extends: the linguistic modality’ (Jackendoff, 2002, p. 349). Similar accounts also gained mainstream support in developmental psychology during the 1970s as the Cognition Hypothesis (Cromer, 1974).

In this account, the process of learning a language consists of matching words to ‘concepts that exist prior to, and independently of, the acquisition of that language’ (Bloom, 2000, p. 242). This view is appealing in its parsimony. As Pinker, Bloom and others have pointed out, it is adaptive to divide the world into categories of objects, actions and properties for non-linguistic reasons. It seems reasonable that words should label these pre-existing non-linguistic categories. This view also has the advantage of enabling a simple, transparent account of compositionality. Words are combined according to grammatical rules; corresponding cognitive rules combine the concepts

these words stand for (Aydede, 2010). For example, when we say ‘the dog runs’, ‘dog’ corresponds to a concept that defines a category of objects, and ‘run’ to a concept that defines a category of actions. By combining these words we express the combination of these concepts. The presumed isomorphism between word and concept is illustrated by the fact that concepts are often represented as capitalised words (e.g. Jackendoff, 2002; Hurford, 2007).³ In this way, ‘the combinatoriality of language serves the purpose of transmitting messages constructed from an equally combinatorial system of thoughts’ (Jackendoff, 2002, p. 272). As Chapter 3 will discuss, this view has its counterpart in evolutionary linguistics – the argument for some form of continuity from animal concepts to word meanings, advanced by Hurford (2007) and others.

However, on closer inspection, there are two main issues with this account. Firstly, the meanings of words vary across languages. Secondly, the meaning of a word varies within a language as a function of context, to the point where, I argue, a view of word meanings as discrete chunks of conceptual structure cannot be maintained. These problems and potential solutions are addressed in the following two sections.

2.2.1 Problem 1: Cross-linguistic variation

If word meanings correspond to non-linguistic concepts that originate from universal human perception of structure in the world, we would expect all languages to have the same repertoire of word meanings. However, this is not the case. Rather, ‘languages display a striking range of crosscutting options for structuring and combining the categories of meanings with which words, grammatical morphemes, and construction patterns are associated’ (Bowerman, 2000, p. 205). This is not only the case for abstract terms in domains far removed from perception, where we might expect the non-linguistic conceptual foundation to be less rigid. It also holds for words that characterise a number of concrete phenomena. For example:

Body parts ‘Two-thirds of the world’s languages have a distinct word for hand. But the remaining one-third does not make this distinction, collapsing hand and arm or hand and lower arm’ (Majid, 2010, p. 65)

³Some researchers have limited patience with this convention: ‘The meaning of a word is represented by the concept WORD, which gets its meaning from OTHER WORDS, which in turn get their meanings from YET MORE WORDS, until it becomes WORDS ALL THE WAY DOWN (they are not words, of course, they are CONCEPTS – hence the capitals – and though no-one knows how they solve these problems, people in this tradition seem happy enough crossing their fingers and hoping that someday, someone will’ (Ramscar, 2010, p. 967).

Containers English speakers label both a wine bottle and a pill bottle as ‘bottle’, whereas Spanish speakers label the former as ‘botella’ and the latter as ‘frasco’ (Malt et al., 1999)

Motion events Spanish tends to conflate Motion with Path in the main verb and express Manner in a participle (e.g., ‘entrar bailando’ = ‘go-in dancing’), whereas English tends to conflate Motion with Manner or Cause in the main verb and express Path in a preposition (e.g., ‘dance in’) (Talmy, 2000)

Spatial relations English uses prepositions to distinguish containment from support, whereas Korean uses verbs to distinguish looseness and tightness of fit (Choi et al., 1999)

Given this cross-linguistic diversity in word meanings, there are three possibilities for the relation between words and non-linguistic concepts:

1. Speakers of different languages also have different non-linguistic concepts. This view is associated with the strong version of the Linguistic Relativity Hypothesis (Whorf, 1956).
2. These differences are not in conceptual content, but only in the mappings between conceptual content and language. In other words, all word meanings are built from the same primitive components; different languages just happen to lexicalise different combinations of primitives (Jackendoff, 2002).
3. Word meanings do not correspond directly to non-linguistic concepts (Langacker, 1976).

I will now consider each of these possibilities in turn.

2.2.1.1 Solution 1: Cross-linguistic variation reflects conceptual variation

The first way to save a direct mapping between words and non-linguistic concepts, while acknowledging cross-linguistic variation in word meanings, is to argue that speakers of different languages have different repertoires of non-linguistic concepts. There are two directions in which this hypothesis could go: a) different cultures encourage the development of different pre-linguistic concepts, which are then reflected in language; b) the language itself constrains the possible concepts that its speakers can entertain.

Most theorists are keen to avoid version b) of the hypothesis, which has historically failed to be supported by data. As summarised in Bloom & Keil (2001), strong claims made based on this hypothesis, for example that Hopi speakers have no concept of time (Whorf, 1956), or that Chinese speakers are less inclined to use counterfactual reasoning (Bloom, 1981), have since been falsified (Malotki, 1983; Au, 1983). In support of version a), some scholars have argued that cultural differences, such as ‘the geographical and interpersonal cohesion of a society’ (Li & Gleitman, 2002, p. 289), can account for differences in conceptualisation that are subsequently reflected in word meanings. However, while this may account for the differences in spatial reference terms discussed by Li & Gleitman, other cross-linguistic differences are harder to motivate. What, for example, are the systematic differences between Korean- and English-speaking cultures that lead the two languages to vary in the ways they express motion events, with English typically conflating Motion with Manner and Korean typically conflating Motion with Path (Choi & Bowerman, 1991), or the differences in spatial relation terms discussed in section 2.2.1 above?

More broadly, several converging lines of evidence show that conceptualisation is at least partly independent from language. Non-linguistic animals such as higher apes have rich conceptual repertoires (for a review, see Hurford, 2007), and children who lack linguistic input, for example deaf children raised by non-signing parents, are still able to form categories at a high level of abstraction, such as generic kinds (Goldin-Meadow et al., 2005). Where language-specific effects on conceptualisation have been demonstrated, they seem to be defeasible, easily induced or suppressed by brief periods of priming (Dolscheid et al., 2013). In the cited study, the authors compared native speakers of Dutch (which uses high/low terminology to describe variation in pitch) to native speakers of Farsi (which uses thin/thick). Participants were asked to sing back tones that were played to them while they were simultaneously shown lines of varying height (height interference condition) or lines of varying thickness (thickness interference condition). Dutch speakers’ pitch accuracy was impaired in the height interference condition, while Farsi speakers’ pitch accuracy was impaired in the thickness interference condition, suggesting that their languages’ mappings between spatial and pitch domains interfered with the task. However, in a followup experiment, Dutch speakers were induced to behave like Farsi speakers by 20 minutes of training in Dutch sentences that used Farsi-like thin/thick descriptions. This result suggests that any effects of language on conceptualisation are defeasible habits, rather than constrained modes of thought. The evidence thus does not support a simple correspondence be-

tween word meanings and non-linguistic concepts.

2.2.1.2 Solution 2: Cross-linguistic variation reflects different combinations of universal primitives

The second possibility is that cross-linguistic diversity is a matter of differences in mappings between parts of conceptual structure and words, rather than differences in conceptual structure itself. This is Jackendoff's position: 'All of these arguments [about cross-linguistic diversity] concern the way elements of linguistic form map into complexes of meaning, not...the contents of meaning itself' (Jackendoff, 2002, p. 292). He argues that lexical concepts in all languages are built up from the same conceptual primitives: different languages just happen to lexicalise different combinations. The motivation behind this approach is to account for a) apparently systematic features of the lexicon, e.g. that the same 'abstract organization' can be seen in many semantic fields (such as parallels between the use of 'be', 'go', and 'keep' for spatial location and possession, Jackendoff, 2002, p. 356), and b) the learnability of word meanings, since 'nearly everyone thinks that learning *anything* consists of constructing it from previously known parts, using previously known means of combination' (Jackendoff, 2002, p. 334; emphasis in original). The argument for lexical decomposition is paralleled in psychology by Eve Clark's semantic feature hypothesis, whereby children build up their word meanings from a universal set of components (Clark, 1973).

The obvious next question is what these conceptual primitives are. This turns out to be less than simple to answer. Jerry Fodor has been a vocal critic of decompositional views of the lexicon, on the basis of two main strands of evidence: 1) working through examples, it is extremely hard to form a definition of any term that does not have counter-examples, as in his analysis of the verb 'paint' (Fodor, 1981); 2) experimental evidence shows that apparently more semantically complex words, e.g. 'kill', often argued to be equivalent to 'cause to die', do not appear to be processed any differently from less complex words, e.g. 'bite' (Fodor et al., 1980).⁴ Jackendoff answers criticism 1) by arguing that we should not look for possible primitives among the meanings of words. Rather, we should be looking for 'layers of structure whose units cannot in-

⁴Although see Gennari & Poeppel (2003), who find increased processing time for eventive verbs (denoting causally structured events) than for stative verbs, which lack this causal structure. However, the problem of completers discussed below remains. Note that Fodor's denial that lexical concepts can be decomposed into primitives, while he still holds to Jackendoff's assertion b), leads him to his position outlined above, that all lexical concepts in all languages must be atomic (unstructured) and innate (Fodor, 1998).

- (12) [_{Sit}GO ([_{Thing}X], [_{Path}P]); [_{Time}T]]
 ('X traverses P over time period T')
 is decomposed as

$$\left[\begin{array}{ccc} [1d]^{\alpha} & [P, 1d_{DIR}]^{\alpha} & [1d_{DIR}]^{\alpha} \\ \parallel & \parallel & \parallel \\ \text{sit}_{BE} ([\text{Thing } X], [\text{Space } 0d]); [\text{Time } 0d] & & \end{array} \right]$$

Figure 2.1: Jackendoff (1996)'s decomposition of GO, previously considered a primitive, into smaller primitives.

dividually serve as possible word meanings...just as we have no conscious access to phonological primitives, we should not be able to expound on word decomposition on the basis of raw intuition' (Jackendoff, 2002, p. 336). However, if we cannot access these primitives by intuition, how can we be sure we are finding the right ones? For Jackendoff himself, this is an ongoing process. In early work he considered GO to be a primitive by which motion verbs could be subdivided (Jackendoff, 1975); however, in more recent work, GO is no longer considered a primitive but is decomposed further (Figure 2.1; Jackendoff, 1996).

Despite Jackendoff's efforts and those of other researchers (e.g. Wierzbicka, 1996), lexical decomposition has not gained a great deal of support. The main issue cited by its opponents is the Problem of Completers: even if we can characterise words like 'run' and 'walk' as sharing the primitive feature GO (or its expanded version in Figure 2.1), there is still a remaining idiosyncratic component to each meaning that remains to be fleshed out (Laurence & Margolis, 1999). In other words, the systematicity of the lexicon is only partial. Jackendoff admits that 'such facts, which we confront at every turn, threaten to undermine the prospect of completely decomposing words into primitives that are descriptively useful and that have some plausibility for innateness' (Jackendoff, 2002, p. 338). Indeed, Eve Clark later abandoned the semantic feature hypothesis for this reason: 'approaches based on semantic components fail overall because only part of the lexicon is compositional' (Clark, 2003, p. 131). Jackendoff's motivation b) for seeking semantic primitives, that without them word meanings would not be learnable, has also been questioned. Schyns et al. (1998) review a number of studies from the psychological concept learning literature that support the idea that the process of category learning involves creating new features, as well as combining pre-

viously known features. For example, Schyns & Rodet (1997) found that varying the order of category learning induced different groups of participants to create different features and to represent and perceive novel stimuli in different ways as a result.⁵

2.2.1.3 Solution 3: Word meanings do not directly correspond to non-linguistic concepts

If words do not correspond directly to non-linguistic concepts, either as atomic wholes or as bundles of universal primitives, the interface between words and conceptual structure may be more indirect. Ronald Langacker argues for this in a 1976 paper: 'Semantic representations, as linguistic objects, are to be distinguished from conceptual structures, the objects of cognition' (Langacker, 1976, p. 322). Versions of this position are taken by Stephen Levinson, Dan Slobin, and Melissa Bowerman, among others (Slobin, 1996; Bowerman, 2000; Levinson, 2003; Evans, 2009). By this account, word meanings are shaped by cognitive constraints and the structure of the world, but also by the specific language being acquired (Malt & Majid, 2013).

Given the intuitive appeal and parsimony of the idea that words label pre-linguistic concepts, this move may seem surprising. However, while conceptual representations are adaptive in principle for both non-linguistic and linguistic functions (recognising and responding to classes of phenomena in the world, and drawing interlocutors' attention to a range of phenomena using a finite vocabulary), the mechanisms by which these two kinds of representation arise are fundamentally different. Non-linguistic concepts are individuals' generalisations over experience; word meanings are conventions, coordinated on in the context of inferential communication and culturally transmitted as new individuals learn them from communicative contexts. This imposes further pressures on word meanings in addition to those at work on non-linguistic conceptualisation. For example, spoken or signed communication is linear in time, requiring events to be sequenced in a way that may not correspond directly to how they are conceptualised non-linguistically (Schouwstra, 2012, p. 109). More generally, if pressures specific to language learning and use shape word meanings without constraining non-linguistic cognition, this account better accommodates cross-linguistic and

⁵In response to critics claiming that these new features could simply be combinations of smaller primitives, Schyns et al. (1998) make a principled argument against fixed feature sets, on the grounds that they will inevitably either be too general or too specific to be consistently useful. 'If the fixed features are fairly high level and directly useful for categorizations...then they will have insufficient flexibility to represent all objects that may be relevant for a new task. If the fixed features are small, subsymbolic fragments...then regularities at the level of functional features, regularities that are required to predict categorizations, will not be captured by these primitives' (Schyns et al., 1998, p. 16)

diachronic variation in word meanings than an account where words map directly onto non-linguistic concepts. Developmentally, this account may also be a better fit with the evidence: reviews of the literature suggest that rather than consisting purely of matching words to pre-existing concepts, a mixture of general cognitive development and language-specific influences are responsible for the learning of word meanings (Schlesinger, 1977; Clark, 2004).

However, the specifics of the indirect relation between words and conceptual structure remain to be fleshed out. If words do not correspond directly to pre-linguistic concepts, how exactly do they relate to our cognition and hence to the world? To clarify this, it helps to move on to the second major problem with the account that words are equivalent to pre-linguistic concepts: the flexibility of word meaning in context.

2.2.2 Problem 2: Contextual variation

Consider some contexts of use of a common verb, ‘run’:

1. John ran down the hill.
2. The dog ran up the stairs.
3. The river runs past the house.
4. The trains aren’t running today.
5. I left the tap running.
6. She runs a bakery.

Based on this evidence, what is the meaning of ‘run’? Most theorists, including those who do not support a straightforward correspondence between words and pre-linguistic concepts, assume that word meaning must nevertheless consist of a stable representation in the language user’s mental lexicon: a discrete chunk of conceptual structure which covers the range of uses of the word, explaining how it is produced and comprehended. Adopting this assumption, this section will lay out two broad possibilities:

- There is one basic or underspecified meaning of ‘run’
- There are a number of distinct meanings of ‘run’

I will argue in the next two subsections that both of these are problematic, and that therefore word meanings cannot be modelled as stable, discrete chunks of conceptual structure.

2.2.2.1 Solution 1: One meaning

The idea of the verb ‘run’ having essentially one meaning is supported by the fact that all the uses above are included in one dictionary entry (OED, 2014). But what does this meaning consist of and how is it fleshed out into the range of uses above? There are two main versions of this idea:

- The meaning of ‘run’ is that expressed in example 1, i.e., the concept of fast human locomotion; the other uses are extensions from this meaning
- The meaning of ‘run’ is an underspecified concept that covers all of its uses

These two accounts fall under the standard two-step model of language comprehension, whereby semantic meaning is first computed on the basis of the literal meanings of the words and their syntactic combination; utterance meaning is then enriched pragmatically by incorporating implicatures, world knowledge, and other extra-linguistic information (Grice, 1975; Sperber & Wilson, 1986; Carston, 2007). This model relies on a clear distinction between semantic meaning (retrieved from the mental lexicon) and pragmatic meaning (inferred from knowledge of the world, the speaker, and the context of the utterance). The discussion below will introduce several converging lines of evidence suggesting that this distinction does not hold.

Basic meaning The intuition that a word has a single ‘basic’ meaning is likely due to a number of factors. Certain uses of a word may be more frequent than others, or more grounded in bodily experience (Lakoff & Johnson, 1980). The basic meaning may therefore be a prototype, in the sense of being the most focal, salient or typical usage of the word, whose contexts of use share the most features with other uses (Rosch & Mervis, 1975; Rosch et al., 1976; Mervis & Rosch, 1981).

However, this does not entail that the basic meaning alone is represented in the individual’s mental lexicon, with the others computed online by extension from this basic use. If the use of ‘run’ in example 1 is basic and the other uses are not stored, just derived from it, we would expect this basic meaning to be accessed on every encounter with the word, whether or not this is the sense that is intended in the context.

However, evidence from priming studies shows that a sufficiently constraining prior context can affect the degree to which the most frequent meaning of a word is accessed, and even whether it is accessed at all (Simpson, 1981; Glucksberg et al., 1986; Paul et al., 1992).⁶ If non-basic uses can be accessed without the basic meaning being activated first, a representation of the basic meaning alone cannot account for the comprehension of polysemous words.

Underspecified meaning Another possibility is that ‘run’ has a single meaning that does not correspond directly to any of the uses above; rather, this meaning is an underspecified concept, RUN, that covers all the uses above in a general way, without encoding features that are specific to particular uses (Caramazza & Grober, 1976; Ruhl, 1989). However, looking at the range of uses of ‘run’ listed above, it is difficult to characterise an underspecified meaning that could generate them all in a principled way. For example, the lack of apparent shared features between use 1 (fast human motion) and use 6 (operate, maintain) leaves it unclear what exactly this representation could consist of: ‘if there is a core [meaning], it has minimal content...As a result, it is not clear what the underspecified meaning could be’ (Klein & Murphy, 2002, p. 566). An underspecified meaning would simultaneously have to be abstract enough to cover all uses, and yet constrained enough to generate a particular range of uses and no others.

Beyond this theoretical issue, there is also experimental evidence against the notion that polysemous words are represented by a single core meaning. Experiment 2 in Klein & Murphy (2001) showed that when participants were presented successively with two occurrences of a polysemous word (e.g., ‘paper’ meaning physical material or newspaper), they were quicker to judge whether the second sentence made sense in the case where the two uses were congruent. In an underspecified meaning account, there should have been no difference, as the same concept would have been accessed in both cases. Croft (1998) also argues that cross-linguistic differences in extension patterns for words that share prototypical uses invalidate a core meaning model. This point will be investigated further at the end of the next section for the case of ‘run’, using examples from English and Greek.

⁶There is some disagreement in this extensive literature about the extent to which context affects or constrains the activation of less dominant meanings; however, a meta-analysis of 25 studies found a consistent effect of context on meaning activation (Lucas, 1999).

2.2.2.2 Solution 2: Several meanings

If the uses of 'run' cannot be accounted for by a single meaning, perhaps 'run' has several meanings. The experimental results cited above against basic or underspecified meaning are usually taken as support for multiple separate meaning representations. However, taking this position leads to further problems. The first is that drawing a principled line between meanings of a polysemous word is extremely difficult and may in fact be impossible. The second is that in doing so, we lose explanatory power by failing to account for relatedness between the meanings, and what this relatedness reveals about the way word meanings are represented and extended.

Boundary between polysemy and vagueness The difficulty in dividing uses of a word into discrete meanings can be seen by looking again at the examples of uses of 'run':

1. John ran down the hill.
2. The dog ran up the stairs.
3. The river runs past the house.
4. The trains aren't running today.
5. I left the tap running.
6. She runs a bakery.

Uses 1 and 2 intuitively seem to constitute the same meaning (although a different number of legs is involved in each case). Use 3 does not involve legs but still involves propulsive motion; similarly with use 4, which in the extension to machinery has also picked up the sense of operating. Use 5 incorporates the flowing liquid feature of use 3 and possibly also the operating feature of use 4. Finally, use 6 has the operating feature without any sense of propulsive motion.

Where should the boundary between meanings be drawn? If it is drawn between 2 and 3, the shared feature of propulsive motion is not accounted for. If drawn between 5 and 6, the shared feature of operating is not accounted for. In short, while some pairs of uses share more features than others, the meanings cannot be divided into principled groups.

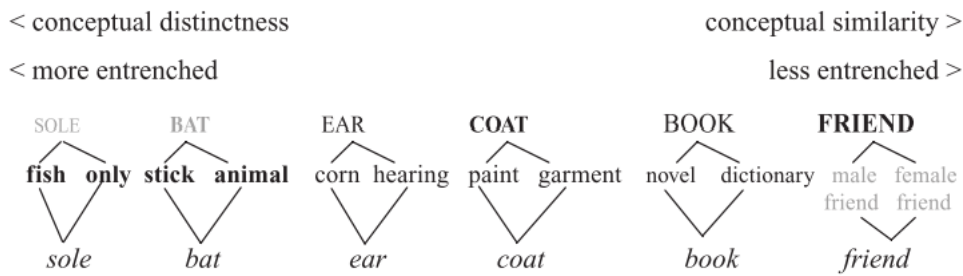


Figure 2.2: Model of homonymy, polysemy and underspecification as regions on a continuum of word reuse. Taken from Murphy (2010), based on a discussion by Tuggy (1993).

Some theorists have argued on this basis that there is no principled difference between a polysemous word and one that is simply vague (as ‘run’ in uses 1 and 2 is vague about whether the running involves two legs or four). Geeraerts (1993) argues for this on the grounds that the tests usually employed to distinguish polysemy from vagueness give inconsistent and contradictory results: ‘For each of the polysemy criteria, examples can be found in which what is a distinct meaning in one context, is reduced to a case of vagueness in another context (and vice versa)’ (Geeraerts, 1993, p. 244). Tuggy (1993) also argues for a continuum between homonymy, polysemy and vagueness, in a model expanded on by Murphy (2010) in Figure 2.2. Here, the difference between polysemy and vagueness is quantitative rather than qualitative: a function of how many features are shared by a word’s contexts of use.

Ruth Millikan argues that the lack of clear difference between polysemy and vagueness is symptomatic of a broader problem with drawing a distinction between what is part of encoded word meaning and what is enriched from context: ‘a clear distinction often cannot be drawn, even in principle, between an expression’s having one sense or several senses, between its being used in a different literal sense or only in an extended or figurative sense, between what has been said and what merely conveyed — hence, more generally, between semantic and pragmatic phenomena’ (Millikan, 2001, p. 405). Taken seriously, this view entails that ‘lexical meanings are not to be thought of as prepackaged chunks of information’ (Geeraerts, 1993, p. 263). This will be further justified in the next section, and a possible alternative explored in section 2.3.

Explanatory power The other issue with classifying a word’s uses under several discrete meanings is a loss of explanatory power regarding how these uses came about. Diachronic and cross-linguistic studies of word meanings show that new uses of a

word can be established by chaining from old uses (Hopper & Traugott, 2003; Hoefler & Smith, 2009). A word is used in a novel context that shares some but not all features of the previous context. On the basis of this contextual similarity, the word assists the hearer in inferring the utterance meaning. This new use of the word is repeated and becomes established as a convention, and in the process, its own features (previously not associated with the word) become available for extension to new uses. The result is a radial or chained family of uses (Lakoff, 1987; Malt et al., 1999; Murphy, 2002), some of which share features (e.g. the bodily motion and liquid motion uses of ‘run’) and some of which do not (e.g. bodily motion and operating).

The partially unpredictable and historically contingent nature of this process can be illuminated by looking cross-linguistically. Words in different languages may share a similar prototypical use, but extend this use in different ways. The Greek verb *τρέχω*, for example, shares the prototypical uses of ‘run’, i.e., human and animal bodily motion (1 and 2). Greek also uses this verb in one of the extended senses used in English, that of a tap running (5).⁷ However, for the remaining examples, the equivalent utterances in Greek would use different verbs. For example, the equivalent of ‘She runs a bakery’ in Greek is expressed with the verb *κρατάω*, whose prototypical use is similar to that of English ‘hold’. Furthermore, Greek makes extensions English does not. For example, two uses of *τρέχω* extend the ‘fast motion’ and ‘effort’ features of prototypical running:

1. Μην τρέχεις όταν γράφεις γιατί θα κάνεις λάθη.
Don’t run when you’re writing because you’ll make mistakes. (*τρέχεις* = rush)
2. Για σας τρέχει όλη μέρα ο πατέρας σας.
Your father runs all day for you. (*τρέχει* = struggles)

For these meaning chains to be possible, and for them to differ cross-linguistically, the possible uses and therefore the meaning of a word cannot be determined only by general cognitive constraints, or by relations between non-linguistic concepts (Croft, 1998). Rather, the patterns of reuse of words must also be shaped by language learning and use. Representationally, too, this means that the traditional view of word meanings as discrete chunks of conceptual structure must be revised. The next section will outline an alternative.

⁷Thanks to Christos Christodouloupoulos for the Greek examples.

2.3 Word meanings as emergent from utterance comprehension

The arguments above have led us to an existentially worrying conclusion. As Dirk Geeraerts puts it, ‘the idea that meanings...do not exist is rather disconcertingly at odds with what we traditionally believe’ (Geeraerts, 1993, p. 259). If ‘one must abandon the concept of word meanings as small discrete chunks of conceptual structure’ (Croft & Cruse, 2004, p. 30), what other model can accommodate the range of uses we observe?

This is not an easy problem. Jeff Elman (2009) considers the issues and proposes the ‘radical surgery’ of doing away with the mental lexicon entirely. Instead, he argues for a model where words function as predictive cues for utterance interpretation, acting directly on the situation models language users incrementally build by integrating these cues with their world knowledge. A word meaning, then, is just the effect that word has on a listener’s situation model. This effect will vary as a function of linguistic and non-linguistic context. Rather than a word linking to a discrete chunk of conceptual structure that generates the range of uses shown above, the uses themselves constitute the word’s meaning. The word’s contribution to the meaning of a given utterance is then shaped by the interaction of the listener’s incrementally built situation model with the history of previous uses of the word. This entails an exemplar-based model, where each new use of a word activates stored instances of its previous contexts of use. The salience of particular exemplars, and particular features of those exemplars, shift as a function of linguistic and non-linguistic context (Figure 2.3). Word meanings are thus conventions in the sense of Millikan (1998): patterns that are reproduced on the basis of specific previous exemplars, forming lineages of uses that are defined by the precedents they stem from, rather than by abstract rules (Millikan, 1998, p. 175).

Jackendoff argues against such ‘contextualist’ models of word meaning on the grounds that they nullify the role of language in utterance interpretation: ‘the expression must convey *something* with which the context can interact. If it did not, a hearer could in principle know from the context what message was intended, without the speaker saying anything at all!’ (Jackendoff, 2002, p. 280). However, if a word does not contribute a discrete chunk of conceptual structure to the meaning of an utterance, this does not mean it contributes nothing at all. What the word contributes is the weighted exemplars of its previous contexts of use, with those weights determined by the context of the current utterance. In this model, word meanings are not privileged

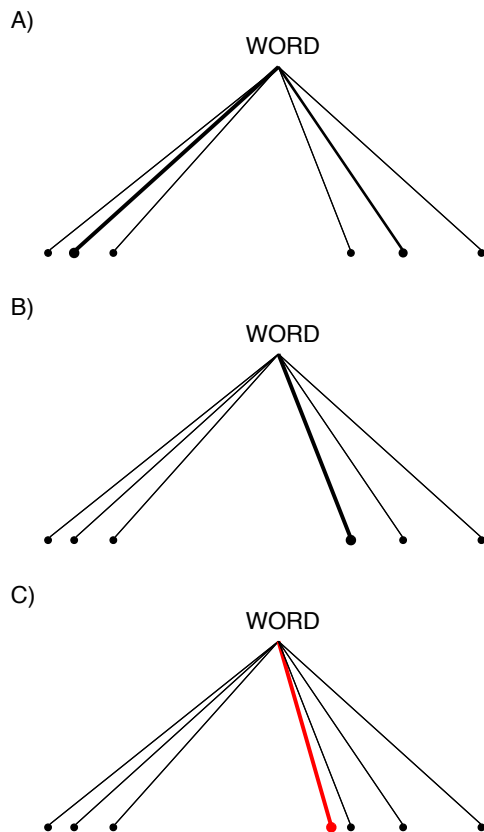


Figure 2.3: Model of word meaning. ‘WORD’ here is a placeholder; this model can also apply to morphemes and phrases. A) Representation of a word meaning in a neutral or default context. The word is associated (lines) with stored exemplars of its contexts of use (dots). These exemplars are embedded in a world knowledge network that includes information about the linguistic (i.e., other words in the utterance) and non-linguistic (e.g., speaker identity, visually salient aspects of the environment) context in which the word was used. Some exemplars have higher default association with the word, determined by frequency, prototypicality, experiential salience, or other factors (bold lines). Speakers may have strong intuitions that these exemplars constitute ‘basic’ meanings of the word. B) Representation of a word meaning in an utterance context. Linguistic and non-linguistic cues selectively weight association strengths of particular exemplars in interaction with world knowledge: e.g., if the word is preceded by another word that strongly predicts a particular exemplar or features associated with it, this will prime that exemplar (bold line). In the case of a novel use, the listener can then infer the word’s relevance in context on the basis of the activated features of that exemplar. C) This novel use of the word, including details of the linguistic and non-linguistic context, is then stored as a new exemplar.

or original: they emerge out of the production and comprehension of utterances. This accounts for the difficulty of objectively defining the meaning units of language, outlined in Chapter 1. If meaning is a property not of words but of utterances, we should not expect it to be consistently resolvable into units. Rather, the exemplars of use of a word or an expression are access nodes into the rich world knowledge network we use for utterance production and comprehension.

Versions of this view are gaining traction among cognitive scientists: see, for example, Spivey (2006); Hagoort & van Berkum (2007); Altmann & Mirković (2009); McClelland (2010); Ramscar (2010); Casasanto & Lupyan (in press). The main motivation for adopting this view is the converging evidence that there is no principled distinction between linguistic and non-linguistic knowledge, or ‘dictionaries and encyclopedias’ (Haiman, 1980, 1982). The consequence is that it is impossible to isolate chunks of conceptual structure that correspond uniquely to word meanings. One of Haiman’s arguments concerns selectional restrictions, or the constraints particular words exert on their arguments (e.g., the fact that the verb ‘eat’ requires an animate agent). These are usually characterised as semantic properties, i.e., part of the meaning of the word (Cruse, 2011, p. 184). However, Haiman extends a point made by Fillmore (1970) to argue that these constraints are in fact part of our world knowledge: ‘The sentence ‘The rock is pregnant’ violates a selectional restriction. But the categorization of rocks as inanimate and hence, a fortiori, barren, is a belief about the world, and one which is not necessarily shared by everyone. (I happen to know of at least one myth, the Hittite story of the monster Uli Kummi, in which a rock does get pregnant.)’ (Haiman, 1980, pp. 345-346).

More recently, psycholinguistic evidence has begun to support this account by showing that there is no principled limit to the knowledge that might be relevant for interpreting utterances in context. For example, a study by Hagoort et al. (2004) found that the assertion that Dutch trains are white (when they are known by Dutch people to be yellow) elicits the same immediate N400 effect as a semantic anomaly (the assertion that Dutch trains are sour). Further, Nieuwland & van Berkum (2006) show that discourse context can even reverse selectional restrictions. In the context of a story about an amorous peanut, animacy-violating assertions (“The peanut was in love”) were processed more easily than those that did not violate animacy (“The peanut was salted”). The immediacy of these effects makes them difficult to accommodate in a model where words correspond to discrete chunks of conceptual structure, with clear boundaries between what is part of word meaning and what is not. Instead, this evidence supports

an incremental model of language comprehension, where listeners integrate linguistic and non-linguistic clues as they become available: ‘The process of comprehension is identical to the process of selecting and verifying conceptual schemata to account for the situation (including its linguistic components) to be understood’ (Rumelhart, 1979, p. 77).

The predictive nature of this process is demonstrated by a large number of studies using the visual world paradigm (Tanenhaus et al., 1995). In these studies, participants’ eye movements are tracked as they listen to utterances while watching a display containing a variety of target objects. Participants’ eye movements show that they anticipate referents before they are named, based on how predictable they are from the words they are currently processing and their general world knowledge (Tanenhaus et al., 1995; Altmann & Kamide, 1999; Chambers et al., 2002; Kamide et al., 2003; Dahan & Tanenhaus, 2004; Brock & Nation, 2014). Factors ranging from selectional restrictions (Altmann & Kamide, 1999) to gender stereotypes (Pyykkönen et al., 2010) are found to have an immediate effect on comprehension. This evidence is not limited to the visual world paradigm; self-paced reading studies show similar effects. For example, the typicality of an agent performing a given action modulates the resolution of syntactic ambiguity (McRae et al., 1997). This account of word meanings as predictive cues for utterance comprehension also explains otherwise puzzling results from the literature on non-literal language comprehension: for example, that speed in processing idioms is determined by their predictability (Cacciari & Tabossi, 1988), or that target words that form part of an idiom are responded to faster in idiomatic (i.e. more predictable) sentences than non-idiomatic control sentences (Estill & Kemper, 1982).

Motivated by these experimental data, this view is broadly compatible with theoretical approaches from cognitive linguistics. Croft & Cruse (2004) argue along these lines that ‘words do not really have meanings, nor do sentences have meanings: meanings are something that we construe, using the properties of linguistic elements as partial clues, alongside non-linguistic knowledge, information available from context, knowledge and conjectures regarding the state of mind of hearers and so on’ (p. 98). Under this account, the patterns of reuse of words do not directly reflect pre-linguistic conceptual structure, but are an emergent product of language’s function as a serial, incremental tool for guessing and communicating intentions.

However, before adopting this account, two potential problems should be addressed. Firstly, how does this model account for the partial systematicity of word meanings?

Secondly, how does it account for compositionality?

2.3.1 Problem 1: Systematicity

This exemplar-based model can still account for the systematic features of word meanings noted by Jackendoff and others. Firstly, the world is partly systematic: categories of phenomena share correlated properties (Mervis & Rosch, 1981) and are organised into hierarchical structures (Tversky, 1989). The lack of a division between world knowledge and linguistic meaning means that patterns of word reuse will partly reflect this structure. Secondly, if words are generalised on the basis of features of their contexts of use, this will lead to groups of words whose uses share features on particular dimensions – the ‘abstract organization’ noted by Jackendoff above. While not completely predictable, this process is still systematic. More broadly, exemplar models do not preclude the possibility of rule-like generalisations: ‘the learning process combines massive storage of examples with the induction of generalizations’ (Hudson, 2007, p. 9). Some examples of higher-order generalisations that language learners may make on the basis of early learned exemplars are discussed in section 3.5.2 of Chapter 3. Geeraerts (1993) fleshes out how prototype effects and the appearance of rule-like generalisations in word meanings must still imply the storage of exemplars, given observed processes of diachronic change:

If a particular peripheral meaning gets to be used more often, and if, at the same time, the originally central meaning gradually lapses into disuse, a conceptual reorganisation takes place resulting in a shift of that application towards the centre of the category. But if the frequency with which an application occurs influences its prototypical status, this implies that there is some mechanism in our conceptual memory for keeping a count (however roughly) of the application’s frequency of occurrence – and this would be the case regardless of the application’s rule-governed nature. If changes in the frequency with which we use an application (or hear it being used) may result in a process by which the application becomes more central in our semantic memory, we are somehow aware of that reading as an individual element in our mental lexicon: the account we apparently keep of the reading’s frequency of occurrence presupposes a representation of that reading — even if it is entirely derivable from a given stored meaning, a particular context, and a specific procedure for semantic extension. (p. 258)

Thus, meaning systematicity can arise as a result of the mechanisms of word reuse in a partially systematic world, rather than having to be pre-specified via the correspondence of word meanings to combinations of conceptual primitives.

2.3.2 Problem 2: Compositionality

In an account where words map directly to non-linguistic concepts, compositionality is simple and transparent. Words contribute their conceptual content to the expressions they appear in, and the combination of these concepts constitutes the meaning of the expression. However, if a word does not contribute a discrete chunk of conceptual structure to an expression, how do we explain compositionality?

The first thing to note is that, as many theorists have pointed out, the classical model of compositionality does not fully account for the phenomena we observe when words combine in natural language (Searle, 1980; Cohen, 1986; Pustejovsky, 1995). The key point is that the contribution a word makes to the meaning of an utterance depends partly on the contributions made by the words it combines with. Searle gives the example of ‘cut the grass’ versus ‘cut the cake’, where the manner of cutting is constrained by the features of the patient and our knowledge of typical events (Searle, 1980). This is a problem for traditional notions of compositionality, since it means a word does not contribute the same content to every utterance it appears in. This then makes the productivity of language a puzzle: ‘If representations changed form from one compound to the next, mastery of a single compound would not entail mastery of another’ (Prinz, 2002, p. 285-6).

The incremental, predictive model sketched above implies a different picture of compositionality. Words contribute to utterance meaning, but in a graded, incremental, and context-dependent manner. Rather than harming productivity, this radical context-dependence may actually maximise the expressivity of language. Cohen (1986) argues that this ‘interactionist’ strategy, where words constrain each other’s possible meanings in context, results in ‘immense gains in [a language’s] ratio of expressive potential to size of vocabulary’: ‘If by their patterns of sentential composition words can impose further subdivisions of sense on one another...or can impose extensions of sense beyond any yet recorded, far fewer words have to be coined and kept in mind for any given set of expressive tasks, and innumerable new tasks can be performed without any need to coin new words. If our own natural language were not already interactionist, it would pay us to make it so’ (p. 226). This notion of meaning flexibility as contributing to the expressivity of language will be explored further in Chapter 3.

2.4 Conclusion

On the basis of the evidence presented in this chapter, I adopt a model where word meanings do not correspond to discrete chunks of conceptual structure that are prior to language. Instead, word meanings are conventions that arise as a by-product of the incremental comprehension of utterances in interaction with world knowledge, and are represented by exemplars of their contexts of use. Following from this, the experiments in the rest of the thesis will use patterns of word reuse as diagnostic of meaning.

This changes the question posed at the beginning of section 2.2. Rather than asking what the concepts associated with words are, the question becomes: what pressures determine the patterns of word reuse we see in languages, and by what mechanisms do these patterns arise and become established? If patterns of word reuse are shaped not only by pre-linguistic conceptualisation, but also by learning and communication, then we can manipulate these variables in artificial language experiments to find out more about the mechanisms by which word meanings arise and become established. Chapter 3 will examine learning and communication more closely in the light of previous work, generating hypotheses about how these pressures might affect patterns of word reuse, and motivating experiments to test these hypotheses.

The evolution of word meanings: Theory, experiments & models

3.1 Introduction

Since Pinker & Bloom's seminal paper 'Natural language and natural selection' (1990), language evolution has grown into a flourishing field (see Fitch, 2010, for a review). Theoretical accounts of the origins of language have been supported by models and experiments exploring how aspects of language could have evolved (Jäger et al., 2009; Scott-Phillips & Kirby, 2010; Kirby et al., 2014). As noted in Chapter 1, much of this work has focused on the origins of syntax and morphology; in comparison, the evolution of semantics has generally received less attention (with major exceptions, e.g. Deacon, 1997; Hurford, 2007).

One reason for this is that many scholars see semantics as evolutionarily straightforward. Pinker & Bloom's view, that a propositional 'language of thought' exists prior to and is expressed by language, represents a common tendency to treat meaning as a solved problem (see Steedman, 2014, for a recent argument along these lines). However, other approaches have questioned this assumption, and modelling and experimental work has uncovered fruitful avenues for investigation of how word meanings arise and develop.

This chapter will first review theoretical accounts of the evolution of word meanings, outlining the theory adopted in this thesis in the light of the descriptive account

proposed in Chapter 2. Following this, I will review previous modelling and experimental work on the origins, communicative emergence and cultural evolution of word meanings. Finally, I will bring these threads together to detail the pressures we might expect to shape word meanings and the mechanisms by which these pressures take effect, motivating a series of experiments to test these hypotheses.

3.2 Theories of word meaning origin

The obvious place to start in a review of theoretical accounts of the origins of word meanings is with Jim Hurford's book *The origins of meaning* (2007). Here, Hurford argues for semantic continuity from our pre-linguistic ancestors to modern humans. The argument, compatible with Pinker & Bloom's, is that semantics exists prior to language: 'Meanings existed in our pre-linguistic ancestors before the application of linguistic labels to them by humans' (Hurford, 2007, p. 57). Hurford bases this argument on the conceptual sophistication of non-human animals, as shown by a variety of observational and experimental studies. This conceptual sophistication increases in species more closely related to humans: for example, while pigeons can generalise over classes of visual stimuli (Watanabe et al., 1995), baboons can hierarchically classify conspecifics by dominance and kin relations (Bergman et al., 2003) and, when trained, can make judgements involving higher-order relations between relations, such as sameness and difference (Fagot et al., 2001). Thus, our closest relatives demonstrate a command of 'abstract relationships that bridge a vast number of disparate behaviors' (Jackendoff, 2002, p.324). Hurford argues that animals' representations of objects and events in their environment (e.g., a lion crouching by a rock) equate to 'proto-propositions' which our pre-linguistic ancestors would also have been capable of representing, and that these proto-propositions are continuous with the meanings of sentences in modern human languages (e.g., 'There's a lion crouching by a rock') (Hurford, 2007, p. 126). More specifically, Hurford argues for direct continuity from an ancestral system of primate calls to words. In his view, 'the English speaker's concept associated with the word *leopard* has a lot in common with what the vervet has in its head that makes it respond systematically to sight of a leopard or sound of a warning bark' (Hurford, 2007, p. 97). While this position is carefully argued for in Hurford (2007), it is also incorporated as an unexamined assumption into other work in the field. Two recent examples are Miyagawa et al. (2014), who argue that 'lexical structure in human language can plausibly be traced to non-human primates and their

alarm calls' (p. 4), and Collier et al. (2014), who characterise Campbell monkeys' 'hok-oo' call as equivalent to the word 'leopard-like' (p. 3).

Beyond the general problems with the notion of word meanings corresponding to 'pre-existing conceptual primitives' (Fitch, 2010, p. 504), discussed in Chapter 2, there are specific problems with arguing for continuity from 'functionally referential' primate calls to human word meanings. Terrence Deacon argues that drawing a parallel between primate calls and words is a 'false lead', springing from our tendency to 'see other species' communications through the filter of language metaphors' (Deacon, 1997, p. 34). A number of critics (Deacon, 1997; Burling, 2007; Rendall et al., 2009; Wheeler & Fischer, 2012) have pointed out important ways in which primate calls differ from words:

1. Their production is largely innate (Owren et al., 1993)
2. They are neurologically more analogous to existing primate call-like vocalisations in humans, such as laughter (Burling, 2007)
3. They are context-specific; unlike words, they do not require integration of context to determine their meaning¹ (Wheeler & Fischer, 2012)

The implication is that primate calls and words are underpinned by fundamentally different mechanisms, suggesting a lack of evolutionary continuity.

Hurford's objection to these arguments is that 'such views leave a gaping hole...They suggest no alternative evolutionary source for modern spoken words' (Hurford, 2012, p.101). And indeed, the problems listed above have not stopped theorists from using primate calls as a way into evolutionary semantics. An important first step is to nail down what these calls actually mean. Hurford sees call meaning as being "word-like": he claims that the leopard alarm call has a meaning similar to the meaning of the word 'leopard', while conceding that for vervets it 'also triggers (or includes?) the typical motor response of running up a tree' (Hurford, 2007, p. 79). Others see the meaning as being more complex: a whole message rather than a word, glossed as something like 'Beware of the [leopard]!' (Wray, 1998, p. 50).² If the first meanings in human language were analogous to the meanings of primate calls, then this disagreement

¹Context may affect production of the calls: for example, vervets are more likely to produce alarm calls when kin are present (Cheney & Seyfarth, 1985). However, call comprehension does not appear to be affected by context.

²The intentional gestures of higher apes are also usually glossed as having a whole-message rather than a word-like meaning: see, for example, Hobaiter & Byrne (2014), where the meanings of chimpanzee gestures are given as 'Stop that', 'Move away', 'Follow me', etc.

leads to two seemingly very different pictures of the origins of linguistic meaning. In the next section, I will outline this debate and argue that it illuminates a further fundamental problem with the notion of continuity from primate calls to modern words. As argued in the previous chapter, words are conventions, coordinated on in communication, learned over cultural transmission, and with their contribution to utterance meaning inferred in a context-sensitive manner. All of these factors, I argue, make primate calls a misleading place to search for the evolutionary origins of word meanings.³

3.2.1 The protolanguage debate

The term ‘protolanguage’ has been used in the language evolution literature to refer to a hypothesised stage between our ancestors having no language, and the advent of modern language as we know it (see K. Smith, 2008b, and the rest of that issue for a range of representative views). While most theorists agree that protolanguage lacked syntax, the semantics of this hypothesised stage have been a matter of fierce debate. However, I will argue that both opposing views build on the same assumption about meaning, and thus are not as different as they first appear.

3.2.1.1 The synthetic account

The synthetic, or lexical, account argues that the meanings of utterances in protolanguage corresponded to modern-day word meanings (Bickerton, 1990; Jackendoff, 2002; Hurford, 2007; Tallerman, 2007). The appeal of this view stems from the assumption that words in modern language (or at least, common nouns and verbs) correspond to atomic concepts which have a long evolutionary history: ‘Nouns and verbs more or less invent themselves, in the sense that the protoconcepts must be in existence before hominids split from the (chimpanzee) genus *Pan*’ (Tallerman, 2007, p. 596). Guy Deutscher also sees this ‘me Tarzan’ stage (Deutscher, 2005) as a plausible starting point for human languages, with proto-nouns and verbs emerging naturally from ‘a conceptual distinction [between objects and actions] that was already there’ (p. 213). According to Deutscher, these elements (plus deictic terms, argued to originate from pointing) are the minimal requirement from which all the complexity of modern lan-

³Arbib et al. (2008) and others argue that the gestures of higher apes may be a better place to look for the evolutionary origins of word meanings, since they show some evidence of being interpreted flexibly in different contexts (Tomasello, 2008). However, these gestures are usually established by ontogenetic ritualisation, rather than being coordinated on as communicative acts, or learned from observation of others. As such, they are still not conventions by the definition above.

guages could emerge by known historical processes of analogy and grammaticalisation; see Heine & Kuteva (2002) for a similar argument.

3.2.1.2 The holistic account

The opposing view, the holistic account, holds that the meanings of utterances in protolanguage corresponded to the meanings of whole sentences in modern language (Wray, 1998; Arbib, 2005; Fitch, 2010). The motivation for this account also comes from continuity with primates, but with the difference that primate calls are interpreted as corresponding to whole propositions, rather than atomic proto-concepts: e.g., glossing the pyow-hack call in putty-nosed monkeys (Arnold & Zuberbühler, 2006), as something like ‘let’s go’ (Zuidema, 2013). Once holistic signals acquire a complex meaning, modern syntactic language emerges by a process of analysis, where learners recognise the recurrence of signal elements in different utterances referring to complex meanings that also share semantic elements (Wray, 1998). This latter prediction has been supported by models and experiments, for example Kirby (2000).

This debate has raged fiercely for the last two decades, with views on both sides ‘pugnaciously defended’ (K. Smith, 2006). I argue that much of the debate arises from misapprehensions about meaning: specifically, the idea that words have fixed meanings that correspond directly to pre-linguistic concepts. Taking into account the arguments against this view presented in Chapter 2 leads to a different picture of the origins of word meanings, outlined in section 3.2.3.

3.2.2 Propositional meaning

Despite their surface differences, the two accounts sketched above share a common assumption: that the meanings of utterances in modern languages – the explananda for evolutionary semantics – are propositions composed out of conceptual atoms. For the synthetic account, proto-words initially corresponded to conceptual atoms; for the holistic account, proto-words initially corresponded to whole propositions composed out of these atoms. The assumed end point for both processes is one where words correspond to conceptual atoms, with syntactic rules determining the way these can be combined into sentences to communicate propositions: i.e., modern language.

The first problem with both of these accounts is that they presuppose a modern semantics. The underlying meanings in each of these hypothesised protolanguages are the same; the only difference is in the allocation of meaning “parts” to parts of utter-

ances. These accounts therefore assume what evolutionary semantics is supposed to explain. Secondly, as theorists from both sides concede, both processes (fractionating complex meanings and concatenating simple meanings) are constantly happening in modern language and can thus be hypothesised to have been at work from the beginning (K. Smith, 2006; A. Smith, 2008a).

Thirdly, and most importantly for this thesis, both of these accounts take a propositional view of meaning, where the meanings of utterances are composed out of discrete conceptual units which are prior to language. Language use is then seen in the following terms: 1) a speaker translates a thought into a string of language; 2) the speaker utters this string; 3) a listener translates the string back into a thought; 4) if speaker and listener now have matching thoughts in their heads, communication has succeeded. In this view, language is a vehicle to ‘represent and share unbounded thoughts’ (Fitch, 2010, p.1). However, this ‘code model’ (Shannon & Weaver, 1949) of communication does not reflect how human language production and comprehension actually works (Sperber & Wilson, 1986; Smith, 2008a). As argued in Chapter 2, meaning is not an a priori property of words; rather, word meanings emerge as listeners integrate the clues provided by words with their guesses of the speaker’s communicative intention in the wider context. The next section will discuss how this view fits into an evolutionary account.

3.2.3 Emergent meaning

What is required evolutionarily to support the emergent model of word meanings proposed in Chapter 2? The answer is surprisingly little. The model requires that language users be highly motivated to predict the communicative intentions of others, adept at inferring these intentions on the basis of limited clues by integrating features of the communicative context with their world knowledge, and skilled at learning how to produce these clues themselves by observing exemplars of their use. The first and second features provide a clue to the human uniqueness of language. Tomasello argues that the crucial adaptation that enabled language is ‘shared intentionality’, or ‘social-cognitive skills for creating with others joint intentions and joint attention (and other forms of conceptual common ground)’ (Tomasello, 2008, p. 11). While other animals have impressive conceptual and learning abilities, humans are uniquely motivated to communicate for the sake of sharing intentions, rather than for purely imperative ends, and, crucially, to interpret behaviour from others in this way (for reviews of the evi-

dence, see Hurford, 2007; Tomasello, 2008). For example, apes in captivity are known to point imperatively, i.e., to request objects or actions, but never declaratively, i.e., merely to share interest, attention, or information (Gómez, 2007) – in stark contrast with human infants, where both kinds of pointing emerge at the same stage of development (Carpenter et al., 1998), even for deaf children raised by hearing parents who therefore lack language input (Goldin-Meadow, 2007). The imperative/requesting tendency in ape communication seems to also hold for gestures used with conspecifics in the wild: all the meaningful gestures observed in Hobaiter & Byrne (2014) are glossed as imperatives, e.g. ‘Move closer’. How and why a more declarative, intention-sharing tendency evolved in humans can only be theorised, but plausible accounts posit it as a consequence of a complex social environment that rewarded greater cooperation and a high general level of interest in other group members’ intentions (Dunbar, 1996; Hurford, 2007; Tomasello, 2008).

Thinking of linguistic communication less as a process of translating thoughts into utterances, and more as a process wherein highly motivated individuals use any clues they can to infer a model of each other’s intentions, the relationship between words and meanings becomes more indirect. To relate this back to the protolanguage debate: under this view, the ‘complexity’ of a given word meaning is hard to define. Words in modern language are usually assumed to correspond to ‘simple’ concepts, but this risks circularity: a concept is simple because it is labelled by a word, and is labelled by a word because it is simple. In fact, even intuitively simple word-labelled actions such as ‘eat’ can be seen as complex by virtue of being composed of a number of sub-events (chewing, swallowing, etc.; A. Smith, 2008a, p. 107). Rather than asking whether the first word meanings were simple or complex, it is more productive to start with the observation that words are learned communicative conventions, and work from there to investigate the pressures that might shape their patterns of reuse.

The next two sections will review previous models and experiments, outlining what this work has revealed about the influence of learning and communication on word meanings. Following this, I will outline the effects we might hypothesise these pressures to have on word meanings, and how I propose to test these hypotheses experimentally in this thesis.

3.3 Models of word meaning

Computational models provide a valuable means of exploring assumptions about the origins and evolution of language, suggesting avenues for experimental work as well as testing theoretical accounts that are experimentally intractable (Jäger et al., 2009; Kirby, 2002). Much of the work described in this section and section 3.4 is concerned with the origins of compositionality: the combination of elements of signals to express a meaning that is a function of the meanings of the signal elements and the way they are combined. As such, these models and experiments typically assume that the world is already divided up into discrete meaning units.

3.3.1 Models with built-in meanings

The models described in this subsection consist of a meaning space, a signal space, and a population of agents. The agents learn by observing signal-meaning pairs (produced either by a previous generation of agents, or in interaction with other members of the population) and updating their associations between signals and meanings according to learning biases. Within this class of models, the iterated learning model (K. Smith et al., 2003b) specifically focuses on intergenerational transmission of signal-meaning pairings. Agents learn from a previous generation of agents, produce utterances based on this learning, and pass their productions on as learning input for the next generation.

A number of models examine the conditions necessary for agents to coordinate on a lexicon (i.e. stable associations between signals and meanings). Agents with learning biases that favour a bi-directional one-to-one mapping between signal and meaning are the most successful in constructing and maintaining a lexicon (Hurford, 1989; Oliphant & Batali, 1997; Smith, 2002). In these models, meanings are built-in to the agents and are transferred along with the signal during communication. This assumes perfect inference of an atomic meaning that is either innate or corresponds to a stable aspect of the world. Luc Steels's Naming Game experiment provides a slight variation (Steels, 1995). Here, the possible meanings are defined by the task: the agents are placed on a virtual map, and must communicate with each other about other agents. To do this, they can either use a word that refers directly to an agent (a proper name-like strategy), or use words that refer to spatial relations (FRONT, RIGHT, SIDE). Over rounds in a communication game, agents converge on a shared vocabulary for communicating these meanings, and agree quickly on new conventions for novel meanings (a new

agent placed on the map). In this model, agents are technically not given built-in meanings; however, agents are able to ‘propagate’ meanings between each other as they come up in conversation before they actually have a vocabulary to express them (Steels, 1995, p. 326). Thus, there is little practical difference between this model and the models above in their treatment of meanings as static and directly transmitted, rather than having to be inferred on the basis of communication. Steels’s later work breaking away from this assumption will be discussed in the next section.

Other models use more complex built-in meaning representations to investigate the origins of compositionality (Hurford, 2000; Kirby, 2001). Early models of this kind used simple predicates with agents and patients as meanings: e.g. LIKE (FIONA, BERTIE). Compositional languages, where meaning elements (i.e. individual predicates and arguments) were expressed by signal elements, emerged as the agents learned from and replaced each other in gradual population turnover.⁴ However, where agents were exposed to a particular propositional meaning more frequently, it tended to be expressed holistically, i.e., with one unanalysed string corresponding to the whole complex event. Whether meaning units are consistently expressed by signal units is therefore partly a function of how frequently those meaning units crop up independently, rather than as part of a frequent complex event.

Later models in this framework use a more abstract feature-based model of meaning.⁵ In these models, meanings are represented as a set of values on a number of feature-dimensions. These meanings could correspond to objects in the world, or equally to conceptual representations in the agents’ minds. The emergence of compositional language in these models is dependent on a number of factors: the size of the bottleneck through which the language must pass at each generation (i.e., the proportion of possible meanings to which learners are exposed, K. Smith et al., 2003b); learning biases that favour one-to-one mappings between meanings and signals (K. Smith, 2003b); and a structured meaning space, where many distinct meanings share values on given feature dimensions (Brighton, 2002). The success of compositional languages in these models results from two factors: 1) these languages respect the learning bias for one-to-one mappings between signals and meanings and 2) they are capable of surviving repeated cultural transmission – i.e., being reconstructible from

⁴The two models differ in that Hurford (2000) assumes the availability of compositional principles to agents, whereas in Kirby (2001), compositionality emerges by virtue of being the most compressible grammar.

⁵Kirby (2000) points out that the agent/patient/predicate model above is mathematically equivalent to a feature-based model.

limited evidence. However, this result is crucially dependent on structure already existing in the meaning space: the same feature values must crop up repeatedly in different meanings in order for the language to survive the transmission bottleneck.

3.3.2 Models with constructed meanings

In the models discussed above, learners induce a language from observing signal-meaning pairs. As Andrew Smith (2003a) points out, ‘the simulations ignore one of the most crucial features of real language acquisition, namely that meanings are *not* transferred with words, and yet learners do manage to infer meanings and associate words with them’ (A. Smith, 2003a, pp. 175-6). To avoid this idealisation, Smith creates a model where feature-values are properties of objects in the world, but there is an additional layer between these objects and the signals used by the agents: ‘a private, agent-specific internal semantic representation’ (A. Smith, 2003a, p. 177). In this model, agents first construct their own semantic representations by creating categories on ‘sensory channels’ which work to discriminate between objects on the basis of their feature-values. These categories then constitute the meanings the agents use to communicate with each other. In a communication episode, the speaker views a target object in a context, finds a meaning which can discriminate the target from the context, and utters a label to express it. The hearer is then given the label and the whole context of objects, and has to infer the intended object based on the associations between the label and the meanings in its internal semantic representation. Communication is considered successful if the hearer infers the correct target object (regardless of whether the agents’ meanings are similar). However, agents are given no feedback on whether communication was successful or not.

This model works as follows: when the hearer sees the context, it searches for every meaning in its own internal semantic representation that would discriminate any one object from all the others. Each of these possible meanings is then associated with the signal. This entails the hearer projecting its own meaning repertoire onto the speaker: an ‘obverter’ strategy (Oliphant & Batali, 1997), which could correspond to an assumption on the part of real-world communicators that their communication partners are similar creatures to them, with similar conceptual repertoires and goals. Agents’ levels of communicative success are therefore highest when their meanings are more similar. Meaning similarity is maximised when two conditions hold: 1) when agents use an ‘intelligent’ meaning creation strategy, i.e., selectively refine the sensory

channels that help most in particular discrimination games; 2) when the structure of the environment is ‘clumpy’. Like the structured worlds of Brighton (2002), this means that a number of objects share values on a particular feature dimension. The effect of this clumpiness is to make particular sensory channels useless for discriminating among members of object groups. As a result of this and the intelligent meaning creation strategy, agents converge on channels that are useful for discrimination, thus boosting meaning similarity. Under these conditions, agents develop highly similar meanings and hence high levels of communicative success.

In this model, meanings do not correspond one-to-one with referents. Rather, a given referent can be conceptualised by a number of different meanings on different sensory channels depending on context (modelling, for example, how you might describe an apple as ‘red’ to discriminate it from an orange, but as ‘round’ to discriminate it from a banana). Kirby (2007) extends this using a model where agents have access to multiple meaning spaces, corresponding to different ways of conceptualising an object. The different meaning-spaces have different numbers of feature dimensions: therefore, in this model, the difference between compositional and holistic languages is a property of meaning spaces, not a property of signals. In several runs of this simulation, multiple meaning spaces remain stable and expressive; however, ‘informative’ meaning spaces (in which more distinctions can be made) win out over ‘uninformative’ meaning spaces (in which few or no distinctions can be made – e.g., a situation where every object is thought of as an undifferentiated ‘thing’). This is a consequence of two factors: 1) the context of communication, such that meanings must be fine-grained enough to distinguish objects from each other, and 2) the transmission bottleneck, such that meanings must be reconstructible on the basis of limited evidence.

A large body of work on the construction of meanings has been done by Luc Steels (for a review, see Steels, 2003). Following on from the Naming Game discussed above, Steels has run a series of studies where robotic agents gradually build up conceptual repertoires that are ‘acquired and aligned in co-evolution with emergent lexicons’ (Steels, 2011, p. 343). In the best-known of these, the Talking Heads experiment, robotically grounded agents ‘not only invent, adopt and align their use of linguistic conventions, but also invent, adopt and align the concepts expressed by these linguistic conventions *based on the outcome of their communicative interactions*’ (Steels, 2011, p. 351; italics in original).

The mechanism of communication is similar to Andrew Smith’s model described above (which was in fact based on the Talking Heads model). The speaker chooses a



Figure 3.1: The setup for Luc Steels' Talking Heads experiment. One of the two embodied robotic agents is chosen as the speaker. The speaker randomly picks one of the coloured figures as the topic and transmits a word to communicate it to the hearer. The hearer has to guess the intended topic on the basis of the word used by the speaker. Figure taken from Steels (2003).

target from a context of objects visible to both speaker and hearer (Figure 3.1). The speaker produces an utterance corresponding to a meaning that discriminates the topic from the context. However, unlike in Smith's model, the meaning that uniquely discriminates the target need not be a category on just one sensory channel. Rather, as in Kirby's model, the meanings can include feature-values on many dimensions (UPPER EXTREME LEFT LOW-REDNESS) or few (LARGE WIDTH). Rather than being defined by any prior notion of simplicity or complexity, the meanings that perpetuate and survive are those that are consistently useful for discrimination. This means that two agents can infer different meanings for a word, but as long as those meanings are confounded in the object sets they have to discriminate, communication will still be successful. An example is 'bozopite', a word that for some agents meant LARGE WIDTH and for some meant LARGE AREA. Only when there were enough discrimination events that involved these features being dissociated (i.e., tall thin objects) did agents converge on a meaning of LARGE WIDTH for this word (Steels et al., 2002). This shows, in the context of an agent-based model, how word meanings can vary across speakers without impacting communicative success, and how the specific patterns of reuse of a given word are a function of the task and the environment, rather than being determined purely by non-linguistic perception. Over time, this led to convergence on meanings that were consistently more useful: a small colour vocabulary and the positions LEFT, RIGHT, UP and DOWN.⁶ These results have since been extended: for example, Vogt (2005) combines the Talking Heads paradigm and the iterated learning model and shows that compositional languages develop on the basis of perceptually grounded semantics, while more recent models demonstrate the establishment of more complex semantic phenomena, such as the emergence of different frames of spatial reference (Spranger & Pauw, 2009).

However, even as these models relax some of the assumptions of the built-in meaning models described in section 3.3.1, they continue to make others. Firstly, in all these models, there is no distinction between linguistic meaning and non-linguistic conceptualisation. This actually works in opposite ways in Smith and Steels's mod-

⁶One caveat to these results is that, unlike in Andrew Smith's model, the Talking Heads agents receive full feedback about the success or failure of their communication game and which object was the intended target. Whether or not feedback is a realistic property to include in experiments and models is debatable; the traditional argument is that language learners do not receive explicit feedback (e.g., Marcus, 1993), but there is evidence against this (e.g., Chouinard & Clark, 2003). The communication games in Experiments 3, 4 and 5, presented in Chapters 4 and 5, provide full feedback to participants, on the grounds that communication failures in the real world lead to implicit feedback by a mismatch between intention and response; however, I acknowledge that this is a simplifying assumption.

els. In Smith's model, agents' meanings are fixed by non-linguistic discrimination games prior to communication, and communication has no effect on these meaning structures. In the Talking Heads model, the discrimination game is embedded in the communication game, so in effect there is no conceptualisation outside of language: 'concepts which have no success in verbal interaction are not encouraged' (Steels et al., 2002). Neither of these seems quite right as a model of how language relates to non-linguistic cognition. On the one hand, language comprehension relies extensively on non-linguistic cognition, and on the other hand, patterns of word use are continually reshaped by communicative processes. Secondly, these more complex models still assume a stable linkage between a signal and a discrete meaning or set of meanings. As argued in Chapter 2, this assumption does not necessarily hold.

To summarise: these models show that the emergence of languages with stable mappings of signals to meaning units rely on a structured world, a bias for one-to-one mappings between words and meanings, and a bottleneck on transmission, i.e., learners being exposed only to a subset of possible word-meaning pairs. In models where meanings are constructed rather than built-in, the meanings that become lexicalised depend on the dimensions that are most useful for discriminating referents from each other.

3.4 Experiments on word meaning

3.4.1 Perceptual meanings

More recently, language evolution researchers have begun building on the results from the models above by running experiments where human participants learn and/or communicate using artificial languages. Kirby et al. (2008) replicate with human participants the model result that compositionality can emerge from an initially holistic language over an iterated learning chain with a transmission bottleneck and a structured meaning space. The meanings in this study are a visual version of the feature-based meanings in the models: images of coloured shapes with arrows indicating their manner of motion, making three dimensions (colour, shape, motion) with three possible feature-values on each. Participants are organised into diffusion chains. As in the model, participants are trained on either a random input language (first generation) or on the language produced by the previous participant in the chain (subsequent gener-

ations).⁷ The participants then produce labels for meanings on the basis of what they have learned, and their productions are passed on as input to the next participant in the chain. In Experiment 2, replicating the model, the languages became compositional: elements of the words began to be reused systematically to express elements of the meanings (i.e., individual feature-values of shape, colour or motion). This result appears to be dependent on a pressure for expressivity (i.e., for the language to continue being able to express all distinctions between all the meanings), either through filtering of duplicate labels (Kirby et al., 2008) or through communication (Kirby et al., submitted).

However, Experiment 1 found a different result that may illuminate a specific property of human learners. Rather than becoming compositional, the languages in Experiment 1 became underspecified: the same word came to refer to a number of meanings that shared a feature-value on one dimension, but differed on others (Kirby et al., 2008, p. 10683). Comparing this to the meaning-spaces model in Kirby (2007), this result can be seen as meaning evolution. Words are reused in patterns that are systematically associated with some features, but not with others. These other features therefore drop out of the meaning space, despite still being perceptible elements of the stimuli. A potential mechanism for this is as follows: 1) via the bottleneck, learners are faced with meanings they do not have a word for; 2) they reuse a word they remember being associated with one of the features of the meaning; 3) over generations, these tendencies build up and are reanalysed as a product of human learners' tendency to condition aspects of language on aspects of meaning (K. Smith & Wonnacott, 2010). For further discussion of the relation of this result to Linda Smith's attentional learning account of the origins of the shape bias, see section 3.5.2 below.

Perfors & Navarro (2014) built on this result using a meaning space where meanings exhibit quasi-continuous variation on two feature dimensions, but one of these dimensions includes a sharp discontinuity (see Figure 3.2). Iterated learning chains were run using each of the meaning spaces shown in Figure 3.2. As the authors predicted, the languages came to lexicalise the distinction between the two sides of the discontinuity, whether size-based or colour-based (right-hand side of Figure 3.2). This shows that the underspecification that arises in these experiments is not random, but is influenced by the structure of the world as well as by human learning biases.

⁷An important point is that all the experiments described here, and those presented in this thesis, use adult participants, not children. The success of the experiments can be interpreted as showing that iterated learning effects do not specifically require child learners, only naive learners. However, known differences between child and adult learners should not be dismissed (Hudson Kam & Newport, 2005).

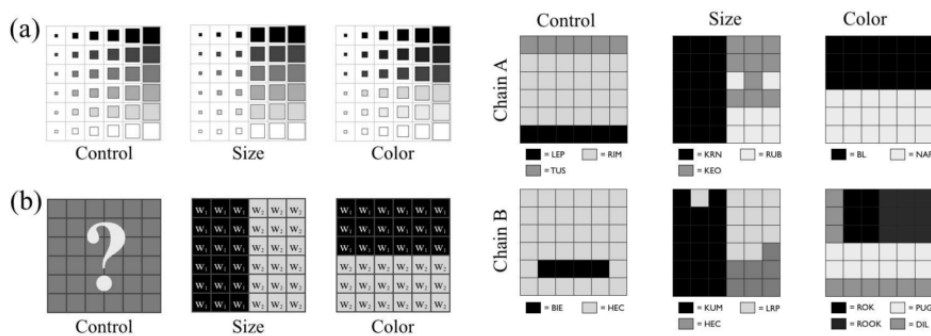


Figure 3.2: Meaning spaces from Perfors & Navarro (2014), showing discontinuous variation in size and colour, plus a control condition where variation was continuous on both dimensions.

To summarise the results from Perfors & Navarro (2014) and Kirby et al. (2008): word meanings (understood as patterns of word reuse) adapt to the learning problem posed by the bottleneck. Patterns of word reuse become more systematically structured, creating word meanings that are more learnable. In this way, the pressures on signals and on meanings are mirror images of each other. If there is a unique signal for each meaning, then signals are hard to learn but the meaning is simple (in the sense that the meaning of a proper noun is simple); if there are fewer signals, the signals are easier to learn but their meanings must become structured in order to remain learnable (i.e., learners acquire a generalisation across features, rather than learning a collection of arbitrary objects). Thus, categorisation emerges as a solution to the learning problem: see also Matthews (2009), where underspecification again emerges in a continuous meaning space, forming categories that exploit features of the stimuli. Overall, these results suggest that iterated learning works to rationalise patterns of word reuse along perceptible dimensions of the stimulus space.

3.4.2 Task-defined meanings

The experiments described above are still somewhat artificial in that they involve overt pairings of signals and meanings in an explicit language learning task. Real language is not like this: we infer meanings as a by-product of doing other tasks, like coordinating action with others. Some experiments have sought to be more realistic by having meanings arise from a task set for the participants, rather than be pre-defined. Two examples of this approach are Galantucci (2005) and Scott-Phillips et al. (2009). Both are coordination games with unusual signalling systems. Galantucci (2005) involves

a number of rooms in a virtual world. A pair of participants view this world from separate computers, each controlling an agent that moves around the space. Their task is to coordinate on moving to the same room. The signalling medium is a scrolling pad that warps any signal traced on it, designed to prevent the participants from writing or using iconic representations. Despite this, participants manage to coordinate on conventions for referring to each of the rooms. Importantly, the rooms are not explicitly presented as meanings to be communicated; rather, they emerge as meanings as a by-product of the coordination task.⁸ This becomes clearer in looking at Game 2 of the experiment, when an additional task is added: agents have to capture a ‘prey’, which they can only do by both being in the same room as the prey. Under the pressure of this new task, many pairs reused their room labels from Game 1 to communicate the new meaning ‘Come to X room, I’ve found the prey’ (with the new feature being inferred from the participants’ shared knowledge of the new task context, rather than explicitly encoded in the signal). However, once this new and salient use of the room signals was established, participants could not use it to simply communicate their location (to avoid a penalty for being in the same room without the prey) for fear of causing costly ‘false alarms’ (Galantucci, 2005, p. 760). This shows how the interaction of world structure and communicative task constrains the possible reuses of old conventions.

A different kind of co-opting of previous conventions is seen in the Embodied Communication Game (Scott-Phillips et al., 2009). This is another coordination game, involving a pair of participants at two computers who each control a stick man in a four-quadrant grid. Their task is to move their stick men to a quadrant of the same colour. Both participants can see each other’s stick man; however, they can only see the colours on their own screen, and the colours on their partner’s screen are different. In this experiment, no signalling medium is provided; the only action participants can take is to move their stick man around the screen. This setup allows investigation of how communicative behaviour is produced and recognised as such in a situation where there is no pre-defined communication channel.

In this experiment, pairs converge on actions (movements of their stick men) that function as labels for colours. However, as in Galantucci (2005), the colours are not pre-defined as meanings per se; rather, these meanings emerge gradually from the task, changing as successive conventions get established. Scott-Phillips outlines a typical trajectory: pairs start by always going to the red quadrant, as red is the most salient

⁸Although note that some participants did orchestrate a ‘naming tour’ of the rooms, showing that explicit labelling strategies are also at play.

colour for coordination. However, at some point, a grid without a red quadrant will come up for one participant. This participant then engages in some marked movement behaviour, communicating something along the lines of ‘do something different’ or ‘not plan A’. The pair then has to pick another colour to move to (e.g., blue). If they succeed in coordinating, the marked movement that previously meant ‘not plan A’ or ‘not red’ comes to mean ‘blue’. The establishment of a lexicon for the game proceeds, with each new convention dependent on the existence of the ones that preceded it, as well as on shared salience biases and the colour-based structure of the game world.

These experiments provide insights how the requirements of a given communicative task, and the conventions already established in a lexicon, work together with the structure of the environment to constrain the possible reuses of existing conventions and the niches that can be taken up by new ones.

3.5 Pressures and mechanisms

Working from the previous results reviewed above and the model of word meaning outlined in Chapter 2, we can now lay out the pressures we expect to be at work on word meanings, and the mechanisms by which these pressures take effect. I will then motivate experiments to test their effects on word meaning structure.

Broadly, in light of the work summarised above, we should expect the following:

- Word meanings should be learnable
- Word meanings should be useful for communication

These pressures are interdependent, since language learning takes place in the ‘arena of use’ (Hurford, 1987). Language learning and communicative inference are therefore closely intertwined. For example, the mutual exclusivity constraint, posited as an innate bias that aids word learning (Markman, 1994), can be explained as a mechanism of communicative inference (Clark, 2007): if a listener hears a novel word in the context of an object for which they already know a label, their assumption of the speaker’s communicative cooperativeness (Grice, 1957) leads them to infer that the novel word must refer to something else. In addition to being interdependent, the pressures of learning and use also depend partly on the structure of the world, since the world is generally what we communicate about. Furthermore, the interaction between the structure of the world and the conventions established so far in the lexicon is also

an important constraint on what patterns of word reuse can develop, as shown by the Galantucci (2005) and Scott-Phillips et al. (2009) results.

This thesis therefore adopts the uniformitarian assumption (Christy, 1983), in terms of uniformity of process: modern processes of language learning and use can be hypothesised to have also applied during the emergence of language. This assumption may not be valid (Newmeyer, 2002). However, given the theory that humans' inferential communicative abilities are a pre-requisite for language (Tomasello, 2008), and the fact that many of our learning and generalisation capacities are shared by higher apes (Hurford, 2007), the next step is to investigate empirically the extent to which these abilities can account for the patterns of word reuse we observe in natural languages.

3.5.1 World structure

Many of the theoretical accounts, models and experiments summarised above make the reasonable assumption that the world⁹ is structured. This is true in a number of ways. Firstly, objects and actions have features that are correlated with each other; for example, the possession of a beak is a reliable cue for the possession of feathers (Mervis & Rosch, 1981). Secondly, objects and actions fall into hierarchically structured categories by virtue of sharing features on particular dimensions; the level at which within-category similarity is maximised and between-category similarity is minimised, the 'basic level', may be particularly cognitively salient (Rosch et al., 1976). Thirdly, objects and actions do not co-occur randomly; instead, particular actions have typically associated participants. This holds for everything from physical possibility, e.g., inanimate objects do not tend to take on agentive roles (Trueswell & Tanenhaus, 1994), to stereotypes, e.g., burly men are more likely than little girls to ride motorcycles (Kamide et al., 2003). Traditionally, the former kind of constraints are classed as selectional restrictions and seen as part of the domain of semantics, whereas the latter are seen as pragmatic properties of general world knowledge. However, as argued in Chapter 2, all of these constraints, on a continuum from absolute to probabilistic, dynamically affect language production and comprehension and hence the development of an emerging lexicon.

To return to a question posed in Chapter 2: if the world is structured, and language is structured, why does language structure not straightforwardly reflect the structure of the world? The work reviewed above shows that learning and communicative pressures

⁹Strictly, the world as filtered through human perception.

interact with world structure, for example in pushing learners to ignore certain perceptible dimensions when reusing words (as in Kirby et al., 2008; Perfors & Navarro, 2014) or in constraining the possible reuses of old conventions (as in Galantucci, 2005; Scott-Phillips et al., 2009). There may, however, be some domains where the structure provided by the world is strong enough that the dimensions which are obvious to perception correlate well with the dimensions specified by words. Colour perception, for example, is an area where it has been suggested that universal cognitive biases amplified by iterated learning straightforwardly explain the distribution of colour categories in the world (Xu et al., 2013). In another example, Dedre Gentner and others argue that objects are cognitively more salient than actions and spatial relations (Gentner & Kurtz, 2005), suggesting that language might map more directly onto their structure without so much interference from learning and communicative pressures.

To summarise: while the structure of the world has an important influence on patterns of word reuse, discrete elements of perception do not always map directly onto discrete elements of language. Learning and communicative pressures interfere between language and the world. The amount of influence that world structure has relative to learning and communication may depend on how salient the perceptible structure in the world is, and on exactly how this maps on to a given learning or communication task. For example, perceptual dimensions that are less salient during learning, or less useful for discrimination during communication, may be less likely to be used as a basis for conditioning patterns of word reuse.

3.5.2 Learning

Word meanings, like words themselves, have to be learnable to survive cultural transmission. Therefore, we should expect pressures from individual and iterated learning to affect the structure of word meanings. This section outlines two key mechanisms by which learning might affect the structure of word meanings.

3.5.2.1 Attentional learning

Humans are very good at learning the meanings of words (Bloom, 2000). Some theorists (e.g., Markman, 1994) argue that this skill is a result of innate biases that constrain the kinds of word meaning hypotheses we are willing to entertain. However, Linda Smith (L. Smith et al., 2010b) presents an alternative account. Rather than having innate biases that, for example, nouns will generalise by shape, we learn these

biases from patterns of word reuse in the lexicon. That is, once a child has (initially slowly) learned a number of words that generalise by shape, she forms a higher-order generalisation that the shape dimension is appropriate to generalise novel words on.

This account has a parallel in evolutionary linguistics, in Terrence Deacon's argument that acquiring a symbol entails grasping a higher-order relation between indices (Deacon, 1997). In Deacon's account, 'a common noun...conveys information about a general type by bringing to mind correlated pairings (indexical uses) with different (iconically related) objects (etc.), and with different combinations of other words (token-token indexicality) in contexts that also share features' (Deacon, 2003, p.121). Like Smith's account, this relies on how adept human learners are at 'recognizing a higher-order regularity in the mess of associations' (Deacon, 1997, p. 89). In other words, the learner observes a number of uses of a word, where a contextual feature on a particular dimension is constant. The learner forms a generalisation about this specific word, that it can be used in contexts that share this constant feature. Once the learner has acquired a number of words that generalise in the same way – i.e., on the basis of a feature on this particular dimension – she can make a higher-order generalisation, that novel words should generalise on the basis of features on this same dimension. As Deacon points out, recognising this kind of higher-order generalisation requires a system that affords making it; he argues that such systems are rare in the natural world and may in fact be unique to culturally transmitted symbolic systems like the lexicon.

3.5.2.2 Iterated learning

As shown by the models and experiments summarised above, iterated learning favours representations that can survive the bottleneck of cultural transmission. What does this imply for meanings? Firstly, iterated learning may work to reduce the number of meaning distinctions in a language. Secondly, the distinctions that remain will be conditioned on particular features of the context, making patterns of word reuse more systematic. The effect of learning interacts with world structure and communication: features that are not shared by many referents, or are consistently unhelpful for discrimination in communicative contexts, are less likely to be the basis for generalisation. Broadly, then, learning leads to a reduction in meaning distinctions, but this reduction is not random: it is directed to particular dimensions that do not harm communication, thus in principle 'enabl[ing] the language user to refer to an unlimited range of specific entities while possessing only a finite number of lexical items' (Pinker & Bloom, 1990,

p. 713).

Summarising, we might expect learning to act in the following ways. On the individual level, attentional learning encourages the formation of higher-order generalisations: e.g., a learner may encounter a set of words that generalise on the basis of shape and therefore generalise novel words on this basis. On the level of cultural transmission, iterated learning works to reduce the number of meaning distinctions in a language and to condition those that remain on particular dimensions, with the choice of these dimensions influenced by the structure of the world and by communication.

3.5.3 Communication

As outlined in section 3.2.2, human communication is fundamentally inferential, involving a recognition of an intention to communicate (Grice, 1957) that may be uniquely human (Tomasello, 2008). Rather than being a process of decoding utterances into thoughts, language comprehension is a process of inferentially constructing meaning on the basis of linguistic and extra-linguistic clues (Sperber & Wilson, 1986). This fact motivates the account in Chapter 2, where words do not have stable meanings.¹⁰ If meanings are instead something we inferentially construct on the basis of patterns of word reuse, then these patterns of reuse should be shaped by the kinds of things we are good at inferring from context. When we reuse a word in a novel context, this new use will only become established if our communication partner can successfully infer the utterance meaning on this basis. This depends in turn on the linguistic and extra-linguistic context of the utterance, incorporating everything from world structure to the specific shared knowledge of the interlocutors.

3.5.3.1 Common ground

Under an inferential account of communication, the possible meanings of a given utterance are bounded only by the mutual knowledge of the interlocutors. This mutual knowledge includes everything ‘from facts about the world, to the way that rational people act in certain situations, to what people typically find salient and interesting’ (Tomasello, 2008, p.75). This mutual knowledge is usually referred to as common

¹⁰Not all theorists who agree that human communication is inferential abandon the code model completely. Sperber & Wilson maintain that words have a core meaning corresponding to a concept in the language of thought (Wilson, 2003), and that words are ‘decoded’ into concepts during language comprehension (Wilson & Sperber, 2004), making language a code nested within an inferential system. However, given the problems laid out in Chapter 2 with the notion of core meanings, I argue that the inferential nature of human communication goes right down to the word level.

ground (Clark, 1996). Tomasello points out that the relationship between communication and common ground is complementary: ‘as more can be assumed to be shared between communicator and recipient, less needs to be overtly expressed’ (Tomasello, 2008, p. 79). He gives an example where the linguistically minimal act of pointing at a rack of bicycles outside a library can work with the rich shared knowledge of the communicators to convey a meaning along the lines of ‘let’s not go in’ (if the pointer knows that a particular bicycle belongs to their companion’s ex-boyfriend and they just broke up last week). Language itself, as a shared system mutually known to the interlocutors, is also part of common ground. Broadly, this account predicts that anything communicators can regularly infer across contexts on the basis of linguistic or non-linguistic evidence (e.g., features redundantly associated with established uses of a word, or features that are consistently predictable from the non-linguistic context) will tend not to be independently lexicalised, or will drop out of the language during the rationalisation process of iterated learning.

3.5.3.2 Expressivity, optimality and alignment

This section will examine three related potential effects of communication on word meanings.

Firstly, a claim often made in evolutionary linguistics is that communication exerts a pressure for expressivity (Pinker & Bloom, 1990; Heine & Kuteva, 2002; Kirby et al., submitted). The general idea is that humans are motivated to communicate all possible distinctions in their meaning space. This pressure therefore works against the pressure from learning to reduce the number of meaning distinctions. In models and experiments, expressivity is generally maintained by avoiding word reuse: agents or participants are biased to pair a single word with a single referent. However, in inferential communication, word reuse is a key strategy for communicating new referents, assuming the context is sufficient for the hearer to infer the novel meaning. As outlined in Chapter 2, repeated episodes of inferential extension result in ‘chaining’ patterns of word reuse, where the individual uses are disambiguated by context. In this way, word reuse can actually lead to greater expressivity without increasing the number of words to be learned (Cohen, 1986; Piantadosi et al., 2012). However, expressivity is still a useful concept to work with, if reformulated slightly: we can hypothesise that communication will favour patterns of word reuse that enable efficient inference of speaker intention, given habitual features of communicative context and the patterns of reuse of words in the rest of the lexicon.

A related argument is that word meanings are in some sense optimised for communication. One expression of this idea is Freyd (1983)'s notion of shareability: 'the structure seen in certain domains of knowledge comes about in the sharing of a set of items' (Freyd, 1983, p. 201). Freyd suggests that the structure of word meanings is determined by mechanisms of analogy that are intrinsic to communication. Gärdenfors (2000), while maintaining that the structure of our concepts partly stems from cognitive constraints, also entertains the possibility that 'communication is a catalyst for geometrically structured meanings' (Gärdenfors, 2000, p.196). As such, we might expect communication to encourage structured patterns of word reuse. However, Freyd makes the caveat that the communicative optimality of a system is partly constrained by its history: the 'shareability' pressure she posits is for novel items to be readily describable by terms already in the system (even if this entails distortion of the features of the novel referent). We might therefore expect communicative patterns of word reuse to be historically constrained by previously established conventions.

A third related claim is that communication leads to alignment between individuals' word meanings. For researchers who do not draw a line between linguistic and non-linguistic conceptualisation, this implies a primary role for communication in the alignment of conceptual structures (Steels & Belpaeme, 2005). There is support for the idea that interlocutors align their representations on many levels during communication (Garrod & Pickering, 2009). While the exact mechanisms are in dispute (e.g., whether alignment is automatic or strategic; whether it is partner-specific or based on priming effects across interactions), the basic result has been shown in a variety of experimental contexts. For example, interlocutors establish and maintain 'conceptual pacts' for referring to objects at particular levels of categorisation (Brennan & Clark, 1996); participants in a maze navigation game align on conceptual schemas for representing positions in the maze (Garrod & Doherty, 1994); and communicating pairs negotiate and align on conventions of reference for novel categories of stimuli (Voiklis & Corter, 2012).

Broadly, then, we might expect communication to encourage structured patterns of word reuse that enable efficient inference of speaker intention across typical linguistic and non-linguistic contexts, and that make efficient use of common ground. We may also expect individuals' patterns of word reuse to align over the course of communication. In general, communication should work against the loss of distinctions encouraged by learning, such that these two pressures balance each other out: for example, there is evidence that kinship categories are a result of a compromise between

simplicity (i.e., learning pressures) and communicative informativity (Kemp & Regier, 2012). However, in making claims about optimality with respect to communication, it is important to bear in mind the mechanisms that may enable or constrain certain progressions of a system. For example, it might be better to stick with a sub-optimal convention and the constrained routes of extension it enforces than to utter something novel and risk misunderstanding.

3.6 Hypotheses

On the basis of the work reviewed above, we can now generate some testable hypotheses about the effects of learning and communication on the structure of word meanings.

Section 3.5.2 outlined the predicted effects of learning on word meanings at the individual and cultural levels. Iterated learning has been shown to result in a) a loss of meaning distinctions and b) the conditioning of the remaining distinctions on particular dimensions of the stimulus space. The attentional learning account argues that individual learners use the patterns of reuse of individual words to form higher-order generalisations about the appropriate dimensions on which to condition their reuse of novel words. However, to make this kind of higher-order generalisation, the learner needs to be exposed to a lexicon that affords it: i.e., a lexicon containing a number of words that generalise on the same dimension. Experiment 1, presented in Chapter 4, tests whether iterated learning can lead to the gradual emergence of a lexicon that affords this kind of higher-order generalisation about the appropriate dimensions on which words should be reused.

Section 3.5.3 predicted that communication should exert a pressure for patterns of word reuse that are expressive, optimised for communication, and aligned within communicating pairs, while acknowledging that these effects might be constrained by early-established conventions. Experiments 2 and 3, presented in Chapter 5, test this hypothesis. Building on these results, Experiment 4, also presented in Chapter 5, adds in cultural transmission, examining how the loss of meaning distinctions imposed by learning interacts with the expressivity, optimality and alignment pressures imposed by communication.

Sections 3.5.1 and 3.5.3 hypothesised that word reuse should be directed to dimensions that are more consistently useful for communication, in terms of being efficient cues to utterance meaning. The efficiency of particular cues is a product of their interaction with common ground (world knowledge and knowledge of the conventions

of a language). Experiment 5, presented in Chapter 6, tests this hypothesis by training participants on event structures in an artificial world where certain features of events are more predictable than others. The experiment investigates whether participants' knowledge of these constraints influences the features that are lexicalised as they develop a language to communicate about the events, and the extent to which early conventions constrain the development of the rest of the system.

3.7 Conclusion

This chapter has presented an evolutionary account of word meanings as emergent from a uniquely human motivation to guess each other's intentions via communication. Word meanings are shaped by the need to be learnable by new generations, and the need to keep language expressive for communication in light of the structure of the world, i.e., the feature distinctions that are salient to human perception. These pressures will push patterns of word reuse in particular directions. Learning will push them to lose distinctions and rationalise those that remain according to features of the context of use. Communication will push towards patterns that allow optimal inference of utterance meanings on the basis of established conventions and world knowledge. The broad picture is of word meanings as a culturally evolved compromise between learnability and expressivity, with the qualification that these pressures are interdependent, and that the mechanisms by which they take effect (feature and event prediction, attentional learning, iterated learning, and communicative inference) constrain the ways in which a system can develop from its first established conventions.

The rest of this thesis will present a series of experiments testing the hypotheses outlined above. The four predictions, to be tested in the experiments presented in Chapters 4-6, are:

- Iterated learning and attentional learning lead gradually to a lexicon that selectively preserves distinctions on more salient dimensions (Experiment 1, Chapter 4)
- Communication leads to qualitatively different patterns of word reuse from individual categorisation (Experiments 2 and 3, Chapter 5)
- Learning and communication work together over cultural transmission to rationalise patterns of word reuse and bring them into alignment (Experiment 4, Chapter 5)

- Patterns of word reuse are conditioned on event features that are less predictable across communicative contexts (Experiment 5, Chapter 6)

The broad question that covers all the experiments is: how do learning and communication interact to influence patterns of word reuse in relation to features of the world? The next three chapters will attempt to shed some light on this question.

The origins of underspecification: Word meanings evolve to selectively preserve distinctions on salient dimensions¹

4.1 Introduction

The previous two chapters outlined the theoretical motivation for the work presented in this thesis. Chapter 2 reviewed descriptions of word meaning from linguistics and psychology, concluding by adopting an exemplar-based model where a word's meaning is the sum of its contexts of use. Under this model, the question of how word meanings arise and develop can be reformulated as follows: what pressures determine the patterns of word reuse we see in languages, and by what mechanisms do these patterns arise and become established? Chapter 3 summarised previous approaches to the origins of word meaning in the language evolution literature and outlined two pressures we might expect to affect word meanings: the fact that word meanings have to be learnable, and the fact they have to be inferrable in the context of communication.

The next three chapters bring these threads together, describing a series of experiments that investigate how word meanings arise and develop over learning and use of artificial languages. The aim is to use these experimental probes to answer broader questions about how pressures of learning and communication work to shape word

¹Part of this chapter is published as Silvey et al. (2014).

meanings, as individuals reuse words in patterns influenced by learning and communicative biases, and cultural transmission amplifies these individual effects. Experiment 1, presented in this chapter, focuses on cultural transmission without communication. Experiments 2 and 3 (Chapter 5) and Experiment 4 (Chapter 6) add communication, as well as exploring the effects of different kinds of perceptual structure in the stimulus space.

Experiment 1 takes as its starting point the observation that words underspecify. Words are reused in contexts which share features on particular dimensions, while features on other dimensions are free to vary. In a traditional account where word meanings correspond to pre-linguistic concepts, this underspecification is a straightforward consequence of conceptual structure. A word might generalise by shape because it labels a concept that defines a category of objects that are the same shape. However, in the account presented in Chapter 2, where word meanings do not correspond to discrete chunks of conceptual structure, we need an alternative mechanism by which these patterns of reuse conditioned on particular dimensions can arise and become established. The results of Experiment 1 show that strong patterns of underspecification can arise gradually over cultural transmission as a product of habitual features of learning and production context. More generally, the work reported in this chapter demonstrates that individual learning effects amplified by cultural transmission can lead to strongly expressed patterns of word reuse that lead new learners to make adaptive generalisations.

4.2 Motivation

4.2.1 Attentional learning and the shape bias

As outlined in Chapter 2, words do not generally have a one-to-one relationship with objects in the world. Instead, language exhibits widespread underspecification: words refer to a range of referents that share features on some dimensions, but differ on others. Different areas of the lexicon have different characteristic patterns of underspecification. Artifact nouns tend to specify shape or function, and underspecify colour; substance nouns tend to specify material, and underspecify shape (L. Smith & Samuelson, 2006). These regularities in the lexicon enable learners to acquire higher-order generalisations about which dimensions are relevant to the meaning of words learned in particular contexts, for example the shape bias that labels for objects generalise by

shape (L. Smith et al., 2002).

Where does this bias come from? One possibility is that shape is simply the best categorisation cue, and that children are aware of this before they begin learning words. However, Linda Smith and colleagues argue that the evidence supports a different hypothesis: the attentional learning account, where ‘the interpretation of a novel word emerges in context from...multiple situational and learned pulls on attention’ (L. Smith et al., 1992, p. 284). For example, for referents labelled by mass nouns, colour or texture are better cues than shape, and children must acquire several nouns of each type before they can consistently generalise newly learned words on the appropriate dimensions. Samuelson & Smith (1999) found that early input to the child is dominated by nouns that generalise on shape (e.g., ‘cookie’) rather than nouns that generalise on material (e.g., ‘applesauce’). The authors suggest that this dominance means there is more data supporting the higher-order generalisation based on shape, explaining the appearance of an early shape bias. In addition, the shape bias can be overridden by the presence of other cues: Jones et al. (1991) found that where referents had eyes, children were more likely to generalise words on the basis of both shape and texture, whereas words for referents without eyes were generalised on shape alone. This suggests that children attend to correlated cues in the context of word learning, building up expectations of what dimensions will be relevant to labelling for particular classes of referent. To achieve this, they need multiple learning experiences with words that specify on particular dimensions. This process is therefore dependent on helpful regularities existing in the lexicon. In other words, the shape bias in the lexicon creates the shape bias in individual learners.

However, this account does not explain how the lexicon comes to have these helpful regularities in the first place. One possibility is that this is a direct consequence of conceptual structure: these underspecified meanings are pre-linguistic concepts, corresponding to ‘words’ in the language of thought (Fodor, 1998; Li & Gleitman, 2002). An alternative account is that these regularities emerge from weak effects of individual learning amplified by cultural transmission. The experiment presented in this chapter tests this second account. Starting from a lexicon which does not preferentially encode distinctions on one dimension over another, a situation akin to the shape bias arises gradually over generations of cultural transmission. The same processes that enable learners to form higher-order generalisations on the basis of regularities in the lexicon can also shape the lexicon to exhibit those regularities in the first place, leading it to reflect the habitual salience of particular dimensions in contexts of learning and use.

This happens not over the course of an individual's learning, but via the cumulative language change that occurs when a lexicon is transmitted.

The attentional learning account states that 'context cues that co-occur with (and define) specific tasks will come with repeated experience to shift attention to the task-relevant information' (L. Smith et al., 2010b, p. 1295). Modelling the learning of (part of) the lexicon as this kind of 'specific task', the experiment has learners trained and tested on an artificial language in contexts where one dimension of meaning is systematically made less salient (backgrounded). I manipulate salience by casting word learning and use as a series of discrimination games where one dimension is never helpful. The general format of the discrimination game has a precedent in the 'guessing game' of Steels (2003), while manipulating one dimension to be unhelpful builds on well-established results in the concepts and categories literature showing that dimensions that are unhelpful for discrimination are attended to less than helpful dimensions (Kruschke, 1992; Medin & Schaffer, 1978). In real word learning, this backgrounding effect is more likely the outcome of factors such as domain-specific knowledge (Kelemen & Bloom, 1994; Lin & Murphy, 1997), increased salience of functional features (Booth & Waxman, 2002; Keil, 1994; Kemler Nelson, 1995), attentional cues from speakers (Tomasello, 2000), inference of the speaker's intention (Bloom, 2000; Xu & Tenenbaum, 2007), or other 'non-linguistic evidence of the speaker's locus of attention' (Clark, 1997, p. 7). In the experiment to be described, this systematic backgrounding has only a small effect at the individual level. However, over cultural transmission, a lexicon that initially specifies equally across all dimensions changes to reflect the differing salience of dimensions in learning and use, leading to a system which preferentially underspecifies the backgrounded dimension. This serves as a demonstration of how cultural transmission amplifies the effects of individual learning processes to create an adaptively specified lexicon, with word meanings that reflect the differing salience of particular dimensions.

4.2.2 Modelling the cultural evolution of underspecification: iterated learning

I model the cultural evolution of language using iterated artificial language learning, as introduced in Chapter 3 (Kirby et al., 2008; Smith & Wonnacott, 2010). Here, I use the diffusion chain instantiation of this paradigm, closely modelled on Kirby et al. (2008).

The meanings in the Kirby et al. (2008) study were a series of images that varied

in shape (square, circle, triangle), colour (black, blue, red) and motion (horizontal, bouncing, spiralling). Each chain was initialised with a language which provided a unique word for each of the 27 meanings: i.e. it specified fully across all dimensions. However, in Experiment 1 of this paper, due to the difficulty of accurately learning and reproducing this language given the amount of training provided, participants began to reuse words for referents that differed on certain dimensions. This led, over several generations of transmission, to the emergence of underspecification as a solution to the learning problem: for example, in one chain, every bouncing square came to be labeled ‘tupim’, regardless of colour.

However, this underspecification was not consistently directed to any particular dimension. Across the different chains, some languages underspecified colour, some shape, and some motion (Cornish, 2011), presumably because, in the learning and testing procedures used in Kirby et al. (2008), no particular dimension was made more or less salient. By contrast, in real word learning and use, some dimensions have higher salience than others (Clark, 1993; Regier, 2005). For particular groups of referents, commonalities across these situations of learning and use will result in certain dimensions being foregrounded and others backgrounded, as per the attentional learning account (L. Smith et al., 2010b). Our hypothesis is that these systematic differences in dimension salience during individuals’ learning and production will lead, over cultural transmission, to a pattern of underspecification that reflects these differences – a helpful lexicon that aids subsequent learners in making the right kinds of generalisations. In order to test this hypothesis, I run a modified version of the Kirby et al. (2008) paradigm, where the learning and production procedures are structured to systematically background one meaning dimension: meanings are presented in pairs that share a feature on one consistent dimension, such that attending to this dimension will never help participants discriminate between the two meanings (Figure 4.1). The hypothesis is that underspecification will gradually arise on the backgrounded dimension, showing that strong constraints on learners’ word meaning hypotheses are not necessary to explain the patterns of underspecification we see in natural language. If, on the other hand, underspecification were to arise indiscriminately on all dimensions, as in Kirby et al. (2008), this would suggest that stronger constraints are needed to explain real-world patterns.

4.3 Method

4.3.1 Participants

40 undergraduate and graduate students at the University of Edinburgh (25 female, median age 20.5) were recruited via mailing lists and organised into 8 diffusion chains. Each chain consisted of an initial participant who was trained on a random language, and 4 successive participants who were trained on the previous participant's test output language, making 5 generations in total: the results of Kirby et al. (2008) suggest that 5 generations would be sufficient for underspecification to arise (in 3 out of their 4 chains the languages had fewer than 5 words by generation 5). Participants in chains 1-6 were unpaid volunteers; participants in chains 7-8 were paid £4.50.²

4.3.2 Stimuli: images and input language

Participants were asked to learn and then produce an 'alien language', consisting of lowercase text labels paired with images. The images were the 27 pictures of coloured shapes in motion from Kirby et al. (2008). The images varied in three possible ways on each of three dimensions of colour, shape and motion (see Figure 4.1 for examples). The training language for the first participant in each chain was a randomly generated set of 27 unique 2-4 syllable labels, built up from 9 possible CV syllables ('da', 'vi', 'ho', 'wi', 'nu', 'ri', 'bi', 'ka', 'tu'). These labels were randomly assigned to the 27 images, ensuring that there was a unique label for every image, with no systematic structure to the labels. The training language for later participants was the language produced by the previous participant in the chain during testing.

4.3.3 Procedure

4.3.3.1 Language learning, language testing, and dimension selection task

The participants worked through a computer program with three phases.

²To ensure that the payment of the last two chains of participants did not affect the results, Chain (i.e. which of the 8 chains of 5 learners a participant belonged to) was modelled as a fixed effect in initial analyses to check if this improved the fit of the models. In all cases, the models including Chain as a fixed effect either did not improve overall fit or showed that no particular chain(s) had a significant effect on the results. In the final analyses below, Chain is modelled as a random effect.

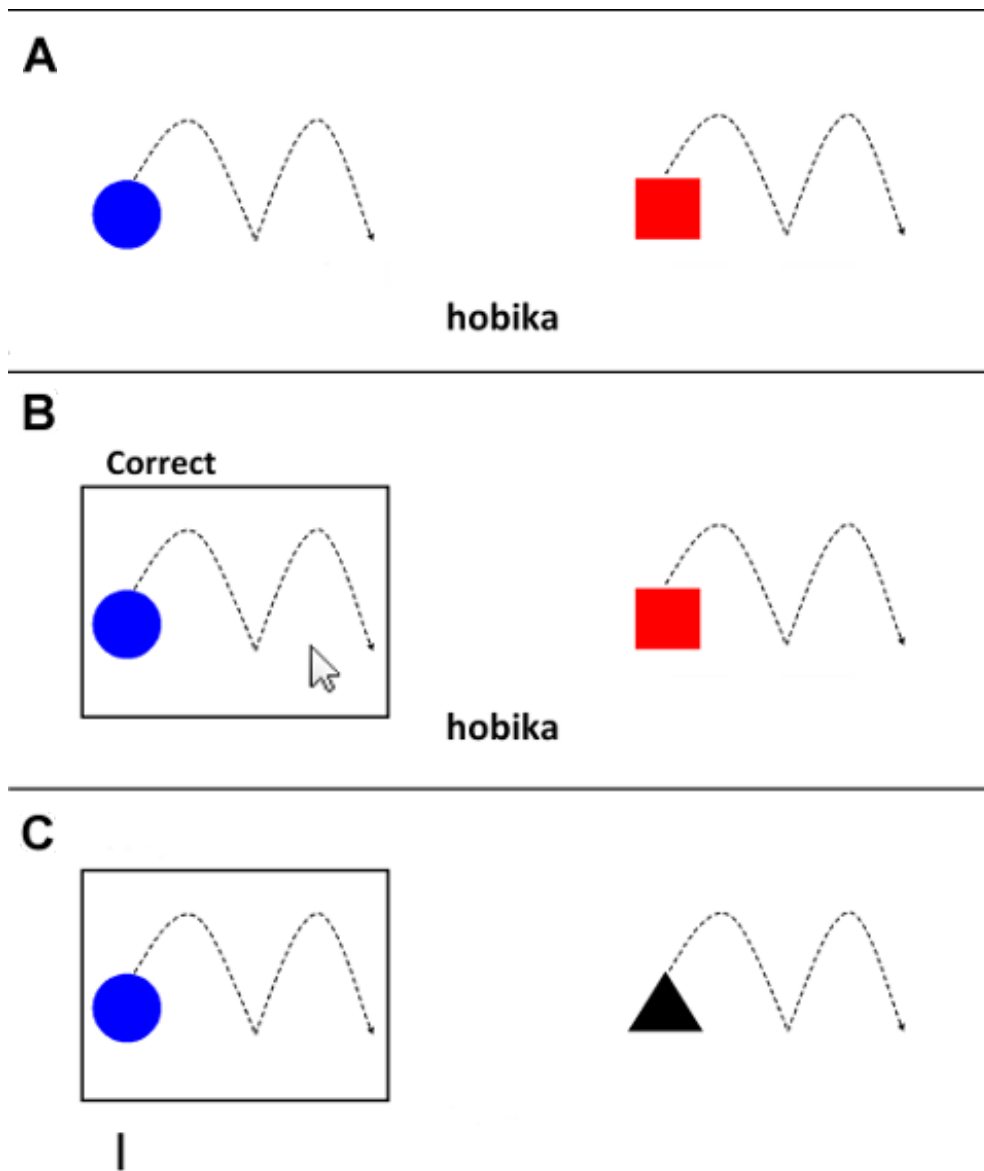


Figure 4.1: Training and testing procedures in the experiment. A) Each training trial is presented as a discrimination game. The participant is shown a word and two candidate images. The participant clicks the image they think goes with the word. B) The participant is then given feedback, followed by the correct word-image pairing. The word then disappears and they are required to retype it. C) Test trials are again presented as a discrimination game, but from the opposite perspective. The participant is presented with two images, one of which is selected as the target. They are instructed to type the word that would allow the alien to pick the correct image. In all training and test trials, target and distractor share a feature on one consistent dimension (in this example, the motion dimension).

Learning phase In each learning trial, the participant was presented with a label and two images, one of which was the target and one a distractor (Figure 4.1A). The participant was instructed to pick which of the two images corresponded to the label. Once the participant had clicked an image they were told whether their choice was correct or incorrect (Figure 4.1B), shown the label and correct image for 2 seconds, and then instructed to retype the label before proceeding to the next trial. Target images were presented in random order. Distractors for each trial of the learning phase were assigned at random, subject to the following constraints: (i) within each learning block, each of the 27 meanings appeared once as a target and once as a distractor; (ii) according to the main experimental manipulation, one dimension was consistently backgrounded during learning and testing trials. For each participant, one of the three dimensions of shape, colour and motion was selected as the backgrounded dimension. Every distractor then had the same feature as the target on this dimension (for example, if colour was selected as the backgrounded dimension, the distractor on every trial would be the same colour as the target). The other two dimensions were not manipulated in this way and served as controls. The learning phase of the experiment consisted of 4 blocks, each of 27 trials.

Test phase In each test trial, the participant was presented with two images: a target and a distractor. The target was highlighted with a black border (Figure 4.1C). The participant was instructed to type the label that would let the alien know which image was highlighted. Target images were presented in random order. Distractors were randomly assigned within the same constraints as in the learning phase, i.e., they matched the target on the backgrounded dimension. The test phase consisted of 27 trials, one for each target.

Dimension selection task This final phase of the experiment used a method from Voiklis & Corter (2012) to test which dimensions participants thought essential to word meaning. On each trial, participants were presented with a label from the language they had been trained on and a concealed image. Their task was to decide whether the label-image pairing was correct or incorrect. In order to do this, they could click to reveal a feature of the concealed image (shape, colour, motion), in any order. Participants could click Correct or Incorrect at any stage and did not have to reveal all features before doing so. A 1-second delay occurred before features were revealed, to discourage participants from revealing features which were unnecessary to make the

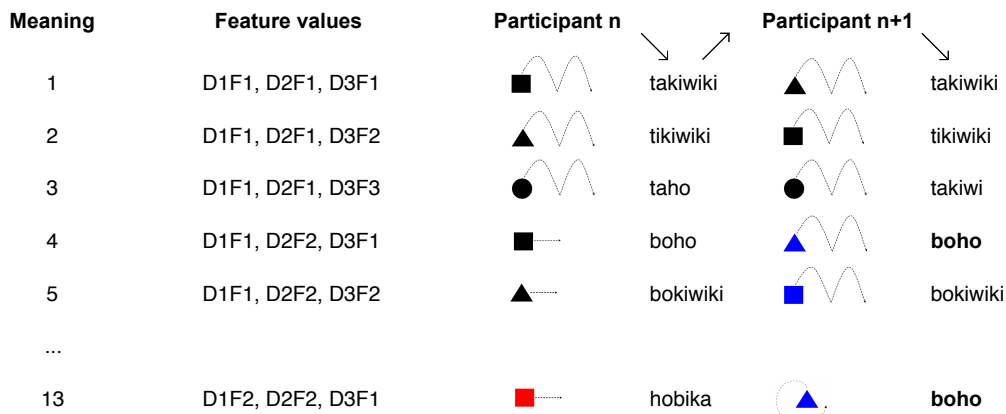


Figure 4.2: Illustration of the transformation process between participants in a chain. In the rest of this legend, 'bk', black; 'rd', red; 'bl', blue; 'ci', circle; 'sq', square; 'tr', triangle; 'ho', moving horizontally; 'bo', bouncing; 'sp', spiraling. During the test phase, participant n produces (downward arrow) 27 labels for 27 meanings, obtained from 3 features on Dimension D1 x 3 features on Dimension D2 x 3 features on Dimension D3. The meaning of each label is represented by specifying the Dimensions D1, D2, and D3 with features F1, F2, and F3 for each dimension. For example, for participant n , D1 is Colour (where F1 = bk, F2 = rd, F3 = bl), D2 is Motion (where F1 = bo, F2 = ho, F3 = sp), and D3 is Shape (where F1 = sq, F2 = tr, F3 = ci). D1 is the backgrounded dimension (here, Colour). The labels produced by participant n are presented to participant $n + 1$ during their training phase (upward arrow); however, the corresponding meanings (i.e., the pictures) are changed randomly, while preserving the backgrounded and salient dimensions. In the example, for participant $n + 1$, D1 is Motion (where F1 = bo, F2 = sp, F3 = ho), D2 is Colour (where F1 = bk, F2 = bl, F3 = rd), and D3 is Shape (where F1 = tr, F2 = sq, F3 = ci). The backgrounded dimension is still D1, but is now Motion rather than Colour. The final column shows the new label produced by participant $n + 1$ during their test phase (downward arrow). Here, we can see that while for participant n 'boho' means black square moving horizontally, and 'hobika' means red square moving horizontally, for participant $n + 1$ 'boho' means blue triangle, regardless of motion. In other words, for meaning 13, this participant produces 'boho' where they were trained on 'hobika', changing the language with this error to introduce underspecification across the backgrounded dimension (motion for this participant).

correct/incorrect judgment. The dimension selection task consisted of 27 trials, one for each image. Images were presented in random order. The labels for each trial were selected from the language the participant was trained on, such that 14 trials contained correct picture-label pairings and 13 incorrect picture-label pairings, but each label appeared only once.

4.3.3.2 Iteration

The language each participant produced in the test phase of the experiment was transformed and then used as the training language for the next participant in their chain. For this transformation, all dimensions and features of the images were randomly shuffled, so that patterns of labeling in relation to backgrounded and salient dimensions were preserved, but individual correspondences of labels to images were not (see Figure 4.2 for an example). This transformation was intended to reduce the effects of intrinsic differences in salience of different dimensions, and to prevent the establishment of iconic labels (e.g., reduplicated syllables for bouncing images).

4.3.4 Dependent variables

4.3.4.1 Transmission error

I used Kirby et al. (2008)'s measure of transmission error (how much the language produced by a participant during testing differed from their training input) to test whether the languages became more learnable over generations. Normalised Levenshtein edit distance between corresponding labels in successive generations (e.g. 'taho' and 'takiwi' for meaning 3 in Figure 4.2) was 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 (see caption to Figure 4.3 for an example of this calculation.) These values were then averaged across the whole language to give one measure of error per participant. If this value decreases over generations, the language is becoming more learnable.

4.3.4.2 Underspecification

Synchronic measures The hypothesis was that the languages would evolve gradually to underspecify more consistently on the backgrounded dimension than on the







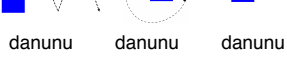
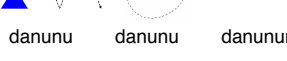

Meaning group	Number of words	Word dissimilarity
 nununu nununu nununu	1	0
 rutunu rutunu rutunu	1	0
 nununu ruhovu hovani	3	0.78
 rinunu rinununu rinununu	2	0.17
 rinunu rutunu rutunu	2	0.22
 rihonu rihova rihova	2	0.22
 danunu danunu danunu	1	0
 danunu danunu danununu	2	0.17
 dahovu dahovu dahovu	1	0
	1.67	0.17

Figure 4.3: Sets of meanings whose labels were compared to obtain underspecification measures (here, with respect to the motion dimension). Meanings were divided into sets of 3 that differed only on one dimension. Word dissimilarity is calculated by averaging normalised Levenshtein edit distances of the three possible word pairs. E.g., for row 4, $\text{rinunu}/\text{rinununu} = 0.25$, $\text{rinunu}/\text{rinununu} = 0.25$, $\text{rinunu}/\text{rinunu} = \text{distance } 0$, so average word dissimilarity is 0.17. (Normalised Levenshtein edit distance for $\text{rinunu}/\text{rinununu}$: 2 letter additions, divided by the length of the longest word (8) = 0.25.) Similar measurements are made over all 9 sets of three meanings that differ only on the motion dimension, and these values averaged to give one underspecification value for that dimension for number of words, and one for word dissimilarity (values in bold).

salient dimensions. Three outcome measures were taken at each generation to assess whether this was happening.

In order to capture the extent to which a language made distinctions on each dimension, I calculated (for each participant's test output) (1) the average number of words the language used across the features on each dimension (possible values ranging between 1 and 3); (2) average normalised Levenshtein edit distance between these labels, to give a more fine-grained measure of label dissimilarity. Figure 4.3 gives an example of how these measures were calculated.

Thirdly, participants' behaviour on the dimension selection task (the order in which they chose to reveal the dimensions) was used to evaluate attention to particular dimensions when evaluating word meaning. I gave a score of 3 for the dimension clicked first, 2 for second, 1 for third, and 0 if the dimension was not selected at all. Dimensions which are selected earlier, and are therefore presumably more central to word meaning, will have higher scores.

Diachronic measure As well as taking the three synchronic measures at each generation, I also looked diachronically at the collapse of lexical distinctions between pairs of images from generation to generation. "Collapse" is defined as follows: in generation n , a given pair of images is referred to by two distinct words, but in generation $n + 1$, this pair of images is referred to by the same word. For every pair of images that differed on only one of the three meaning dimensions, I coded whether this dimension was backgrounded or salient, and in which generation (if at all) the distinction between the two images collapsed. This gives us a dynamic view of the changes in underspecification in the languages over generations.

4.3.4.3 Language structure

I used a measure from Kirby et al. (2008) to test whether the languages became more structured over generations. The aim was to quantify how systematic the mapping from words to images was across the whole language. If words that are similar to each other map onto images that are similar to each other, this demonstrates that the language is systematically structured. To measure structure, I took a) the normalised Levenshtein edit distances between all pairs of words in each language and b) the Hamming distances between the corresponding pairs of images (in terms of shared features: e.g., a bouncing red circle and a bouncing red triangle have different features on one dimension, shape, and so have a distance of 1). I then calculated the correlation between

these two distance matrices. To assess the extent to which this veridical correlation differed from what would be expected by chance, I ran a Monte Carlo simulation to shuffle the matrix of meaning distances 10000 times, recomputing the correlation each time and adding it to a distribution of values for this language. The final structure measure is the z -score for the veridical correlation based on the distribution of shuffled values.

Building on Kirby et al. (2008)'s measure to apply it to the underspecification hypothesis, I calculated three versions of the structure measure, each calculating meaning distance in a different way. The basic structure measure coded Hamming distance using all three meaning dimensions: i.e., if two meanings differed by one feature on any dimension, they had a distance of 1. The underspecified structure measure coded Hamming distance ignoring the backgrounded dimension: i.e., if two meanings differed only on the backgrounded dimension, their meaning distance was counted as 0. Finally, a control version of the structure measure calculated Hamming distance ignoring each of the salient dimensions in turn, and then averaged the two structure measures produced. The hypothesis predicts that structure emerging in the language will preferentially apply to the two more salient dimensions; therefore, the underspecified structure measure should, over generations, become higher than the basic structure measure, and the control measure (ignoring one or other of the salient dimensions) should become lower than the basic structure measure.

4.4 Results

4.4.1 Transmission error

Transmission error decreased over generations (Figure 4.4). Error in generation 1 was 0.67, 95% CI [0.60, 0.74], decreasing by generation 5 to 0.34 [0.20, 0.48]. This was a decrease of 0.33 [0.19, 0.47]. A linear trend ANOVA found that the trend over generations was significant, $F(1,7) = 27.84$, $p < .001$, showing that the languages changed to become more learnable. Cohen's unbiased d (hereafter d_{unb}) was calculated for this effect by dividing by the pooled SD for error at Generations 1 and 5, and then multiplying by Hedges' adjustment factor, $1 - (3/(4df - 1))$. d_{unb} was 2.34, suggesting a very large effect.

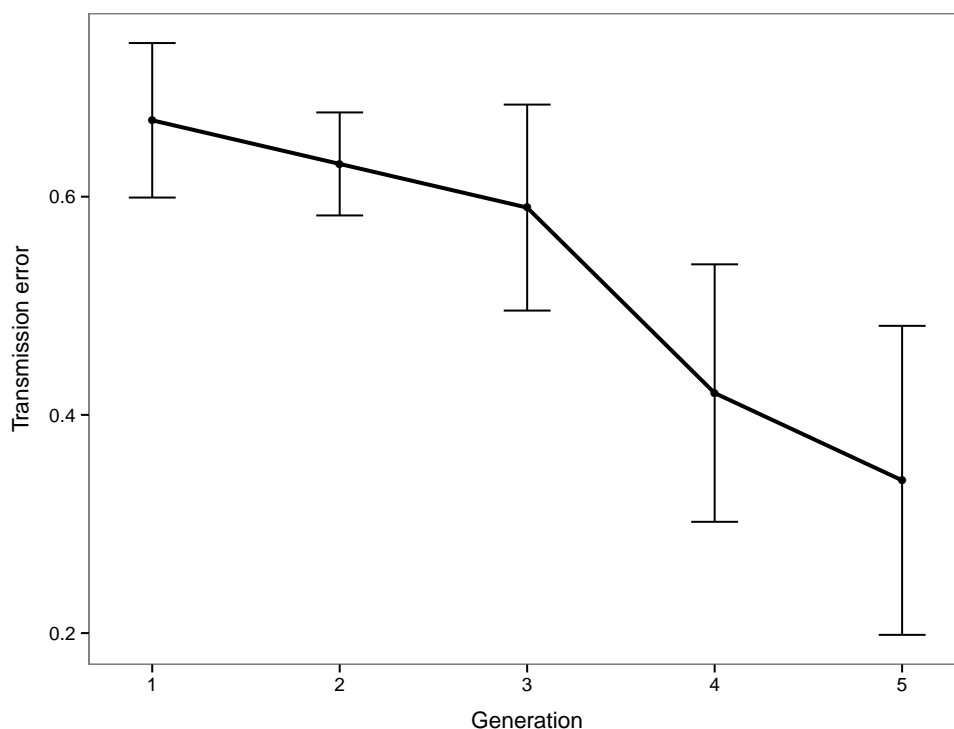


Figure 4.4: Transmission error over generations (see section 4.3.4.1 for how this is calculated). Error bars show 95% confidence intervals.

4.4.2 Underspecification

The results for the two linguistic measures of underspecification, within-dimension number of words and label dissimilarity, are shown in Figures 4.5 and 4.6 respectively, with the result for the dimension selection task in Figure 4.7. Mixed-effects models were used for the main analyses of each of these dependent variables (number of words, within-dimension label dissimilarity, dimension selection task). The random effects to include in these models were assessed by means of likelihood ratio tests. All models incorporated a random intercept for Chain and a random slope for Participant by Dimension (salient/backgrounded). p -values for the fixed effects in these models were estimated using Baayen (2008)'s formula.³

For post-hoc tests, the observations for the two salient dimensions were averaged. t -tests were then run comparing backgrounded and salient dimensions at each genera-

³ $2 * (1 - pt(abs(t), Y - Z))$, where Y is the number of observations, and Z is the number of fixed effect parameters. The pt command on R accesses the probability distribution for t . $Y - Z$ calculates the degrees of freedom, and multiplying by 2 obtains the p -value for a two-tailed test. Since this can be anticonservative at small sample sizes, I also used the heuristic of only accepting t values larger than 2 as significant (Baayen, 2008).

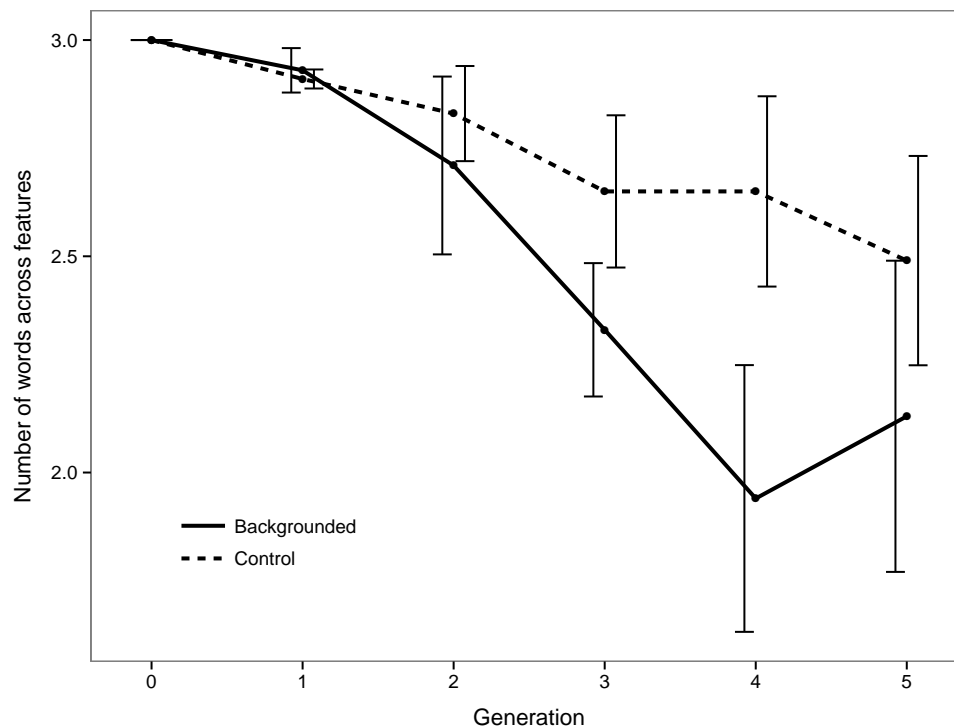


Figure 4.5: Number of words across features (see Fig. 3 for how this is calculated) against generation. The solid line indicates the backgrounded dimension, while the dashed line shows the salient dimensions. Error bars (offset for clarity) show 95% confidence intervals, with standard errors adjusted to reflect only between-subjects differences.

tion, applying the Bonferroni correction for multiple comparisons.

4.4.2.1 Synchronic measures

Number of words Figure 4.5 shows the change in number of words across features over generations. Mean number of words across backgrounded and salient dimensions was similar at generation 1: 2.93, 95% CI [2.84, 3.02] on the backgrounded dimension and 2.91 [2.84, 2.98] on salient dimensions, a difference of 0.02 [-0.05, 0.09].

The number of words across backgrounded and salient dimensions then gradually diverged over generations 2-5, with more words remaining on salient dimensions than on the backgrounded dimension. The greatest difference was in generation 4: 1.94 [1.61, 2.27] on backgrounded dimensions and 2.65 [2.37, 2.93] on salient dimensions, a difference of 0.71 [0.31, 1.11]. Fixed effects of dimension salience, generation, and an interaction were included in the mixed-effects model. Analysis of this model

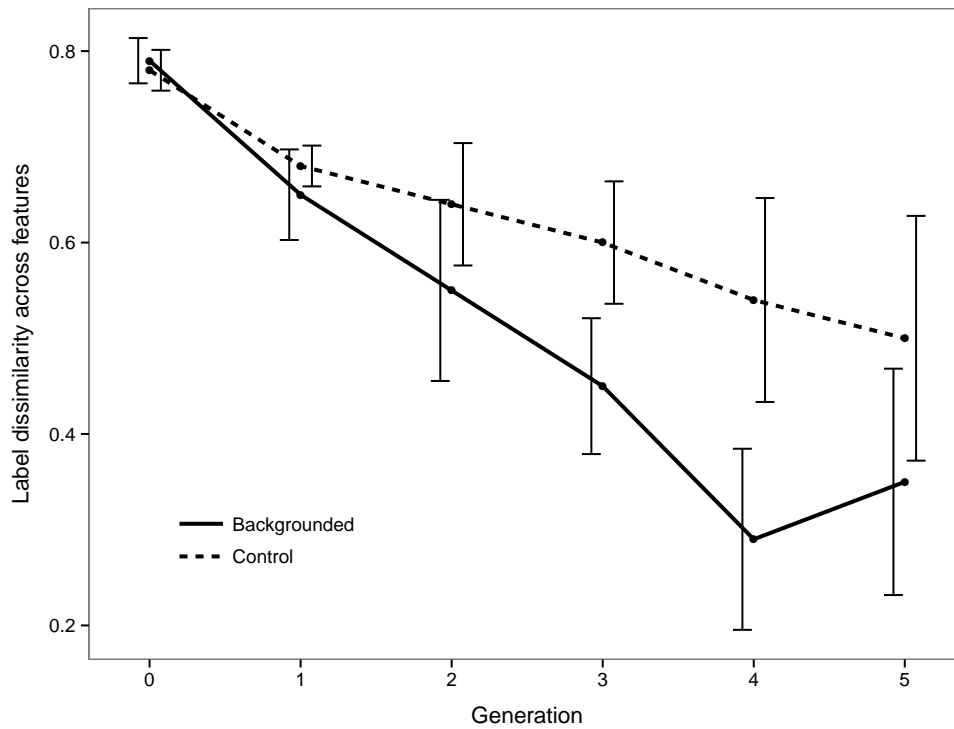


Figure 4.6: Dissimilarity of labels across features (see Figure 4.3 for how this is calculated) against generation. The solid line indicates the backgrounded dimension, while the dashed line shows the salient dimensions. Error bars (offset for clarity) show 95% confidence intervals, with standard errors adjusted to reflect only between-subjects differences.

showed that the main effect of dimension salience was significant, $\beta = 0.25$, $SE = 0.06$, $t(144) = 4.46$, $p < .001$. There was also a significant linear trend for the number of words to decrease over generations, $\beta = -0.92$, $SE = 0.11$, $t(144) = -8.29$, $p < .001$, and the effect of generation was also significantly different for backgrounded versus salient dimensions, $\beta = 0.50$, $SE = 0.14$, $t(144) = 3.66$, $p < .001$.

Post-hoc tests (using the Bonferroni correction to establish a significance criterion of .008) found that the difference between backgrounded and salient dimensions was marginally significant in generation 3, $t(7) = 3.54$, $p = .009$, and significant in generation 4, $t(7) = 4.03$, $p = .005$. d_{unb} was calculated for this within-subjects difference by dividing by the average SD for backgrounded and salient dimensions, and then multiplying by Hedges' adjustment factor, $1 - (3/(4df - 1))$. d_{unb} for these differences was 0.97 in generation 3 and 1.45 in generation 4, suggesting a very large effect. The difference was not significant in any other generation, $t(7) < 1.71$, $p > .13$.

Within-dimension label dissimilarity Figure 4.6 shows the change in within-dimension label dissimilarity over generations. Mean Levenshtein edit distance between words across backgrounded and salient dimensions was similar at generation 1: 0.65, 95% CI [0.58, 0.72] on the backgrounded dimension, 0.68 [0.63, 0.73] on salient dimensions, equating to a difference of 0.03 [-0.04, 0.10]. These values then gradually diverged over generations 2-5. Words became more similar (i.e., edit distance was lower) on the backgrounded dimension than on salient dimensions. The largest difference was in generation 4: mean Levenshtein edit distance between words on the backgrounded dimension was 0.29 [0.15, 0.43], compared to 0.54 [0.42, 0.66] for salient dimensions – a difference of 0.25 [0.11, 0.39].

The mixed-effects model incorporated main effects of dimension salience and generation, plus an interaction. Analysis of this model showed that the main effect of dimension salience was significant overall, $\beta = 0.11$, $SE = 0.03$, $t(144) = 4.30$, $p < .001$. Additionally, word dissimilarity tended to decrease over generations, $\beta = -0.40$, $SE = 0.05$, $t(144) = -7.86$, $p < .001$, and the effect of generation was also significantly different for backgrounded versus salient dimensions, $\beta = 0.19$, $SE = 0.06$, $t(144) = 2.97$, $p = .004$.

Post-hoc tests (using the Bonferroni correction to establish a significance criterion of .008) found that the difference between backgrounded and salient dimensions was significant in generations 3, $t(7) = 3.92$, $p = .006$, and 4, $t(7) = 4.42$, $p = .003$. d_{umb} for this difference was 1.01 in generation 3 and 1.23 in generation 4, suggesting a very large effect. The difference was not significant in any other generation, $t(7) < 1.95$, $p > .09$.

Dimension selection task Figure 4.7 shows the change in behaviour on the dimension selection task over generations. Mean selection preference score for backgrounded and salient dimensions was similar at generation 1: 1.72, 95% CI [1.53, 1.91] for the backgrounded dimension, 1.78 [1.66, 1.90] for salient dimensions, equating to a difference of 0.06 [-0.18, 0.30]. This pattern was similar in generation 2, then gradually diverged over generations 3-5, with higher preference scores for salient dimensions than for the backgrounded dimension. The difference was largest in generation 4: 0.98 [0.41, 1.55] for the backgrounded dimension compared to 1.89 [1.51, 2.27] for salient dimensions, equating to a difference of 0.91 [0.67, 1.15].

The mixed-effects model included fixed effects of dimension salience, generation, and an interaction. This model found significant main effects of dimension salience

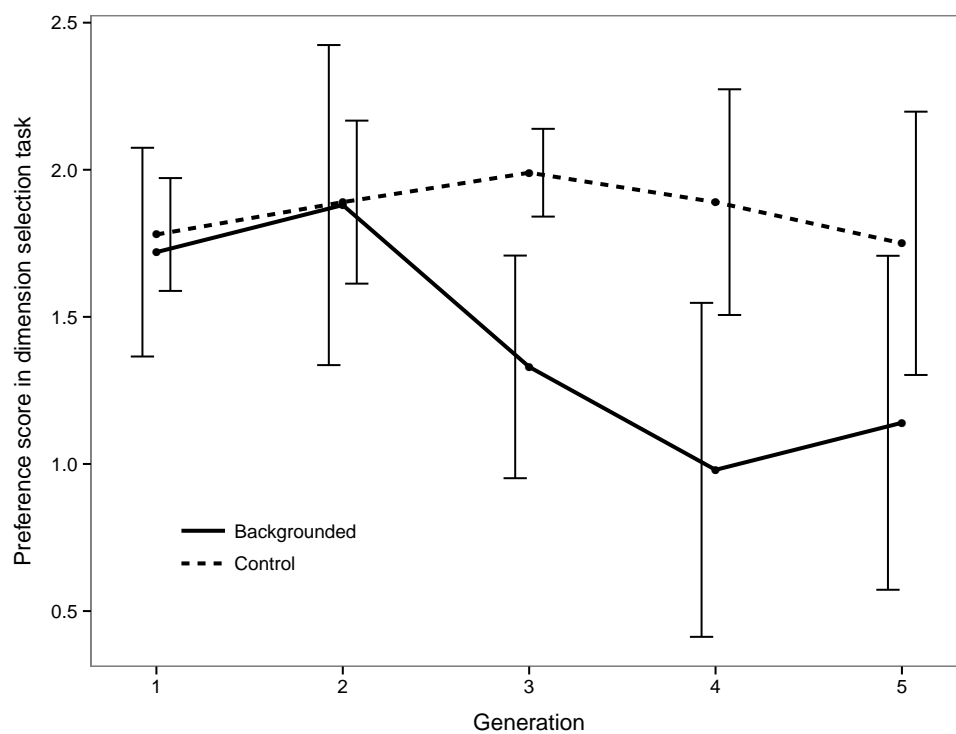


Figure 4.7: Change in attention to different dimensions over generations, evaluated via the dimension selection task. The solid line indicates the backgrounded dimension, while the dashed line indicates the salient dimensions. Error bars are 95% confidence intervals.

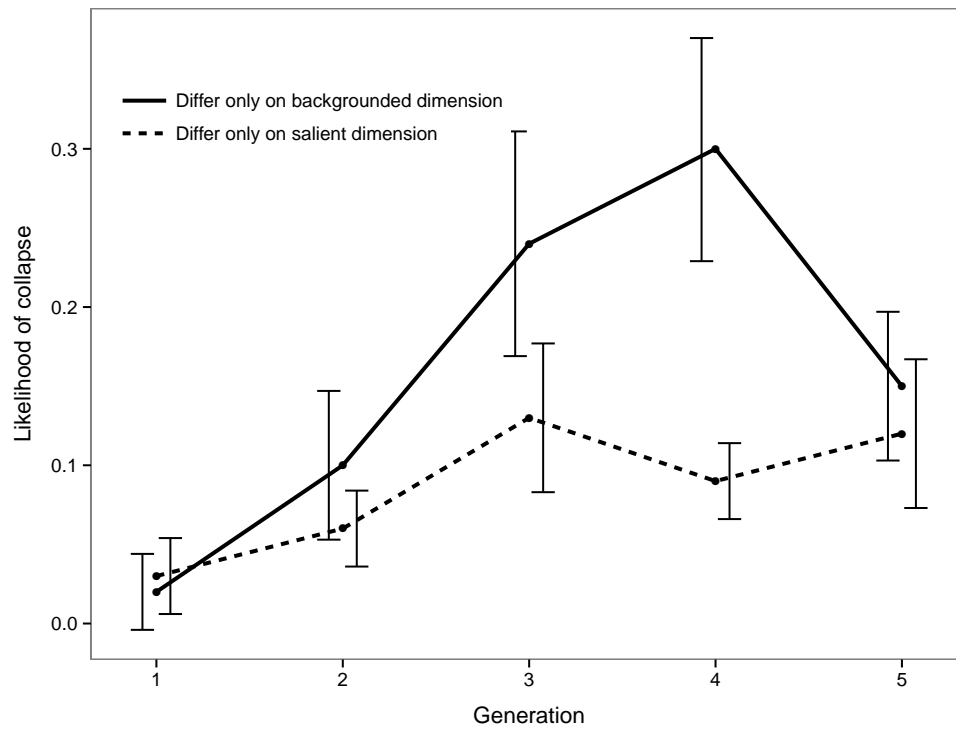


Figure 4.8: Likelihood of collapse of distinctions at each generation between pairs of images that differed only on one dimension. The x-axis shows output generation - i.e., 1 stands for the transition from generation 0 to generation 1. The solid line shows the likelihood of collapse for pairs that differed only on the backgrounded dimension, and the dotted line the likelihood of collapse for pairs that differed only on a salient dimension. Error bars are 95% confidence intervals, with degrees of freedom based on the number of chains (i.e. $df = 7$).

($\beta = 0.45$, $SE = 0.12$, $t(3240) = 3.85$, $p < .001$), generation ($\beta = -0.65$, $SE = 0.21$, $t(3240) = -3.14$, $p = .002$), and a significant interaction between the two ($\beta = 0.63$, $SE = 0.26$, $t(3240) = 2.42$, $p = .016$).

Post-hoc tests found a significant difference between backgrounded and salient dimensions only in generations 3, $t(7) = 3.92$, $p = .006$, and 4, $t(7) = 3.70$, $p = .008$ (significance criterion using Bonferroni correction = .01). d_{umb} for this difference was 0.55 in generation 3 and 0.78 in generation 4, suggesting a medium to large effect. The difference was not significant at any other generation, $t(7) < 1.09$, $p > .11$.

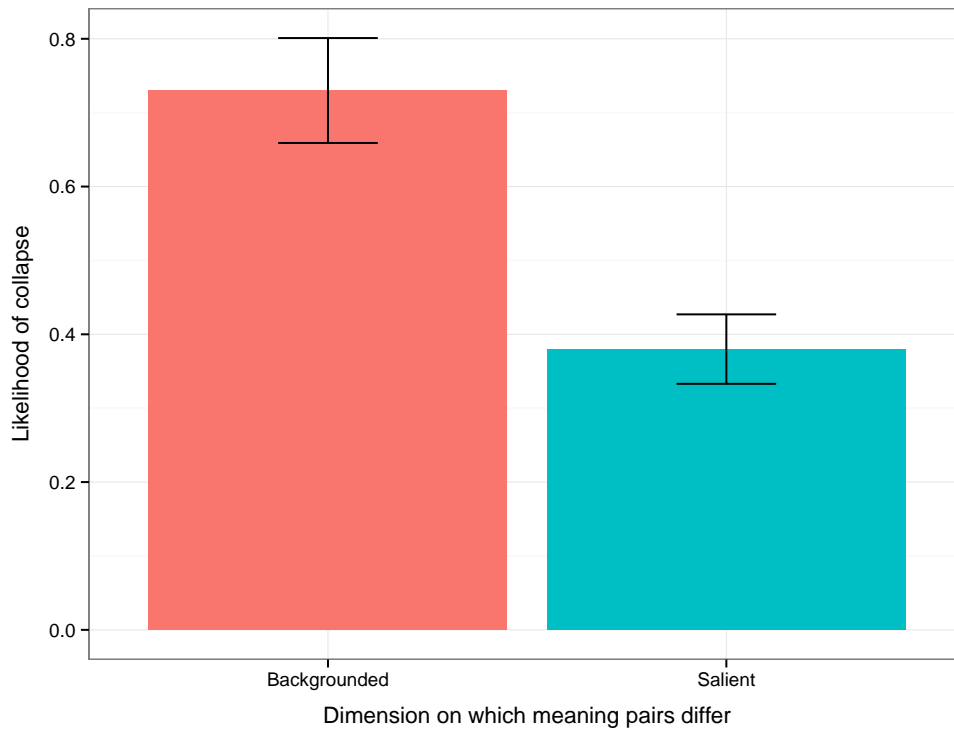


Figure 4.9: Likelihood of collapse of distinctions between pairs of images that differed only on one dimension. The red bar shows the likelihood of collapse for pairs that differed only on the backgrounded dimension, and the blue bar the likelihood of collapse for pairs that differed only on a salient dimension. Error bars are 95% confidence intervals, calculated on the basis of the number of chains (i.e. $df = 7$).

4.4.2.2 Diachronic measure

Figure 4.8 shows the likelihood of collapse between each successive pair of generations for pairs of images that differed on only one dimension. Collapse can only happen once over the chain for each pair of images, unless a distinction between them is regained and then lost again. This affects the degrees of freedom, making these results difficult to analyse statistically and constraining the possible trends. For example, if a distinction has collapsed in generation n , it cannot collapse in generation $n + 1$. The low figures for collapse in generation 5 are therefore partly a consequence of the fact that so many distinctions have already collapsed in generation 4 (although see section 4.5.4.2 for further discussion of the generation 5 results).

To solve this problem, I pooled the data over generations and coded each pair of images as either collapsing at some point during the chain, or not collapsing at any point during the chain. The results for pairs that differed only on the backgrounded

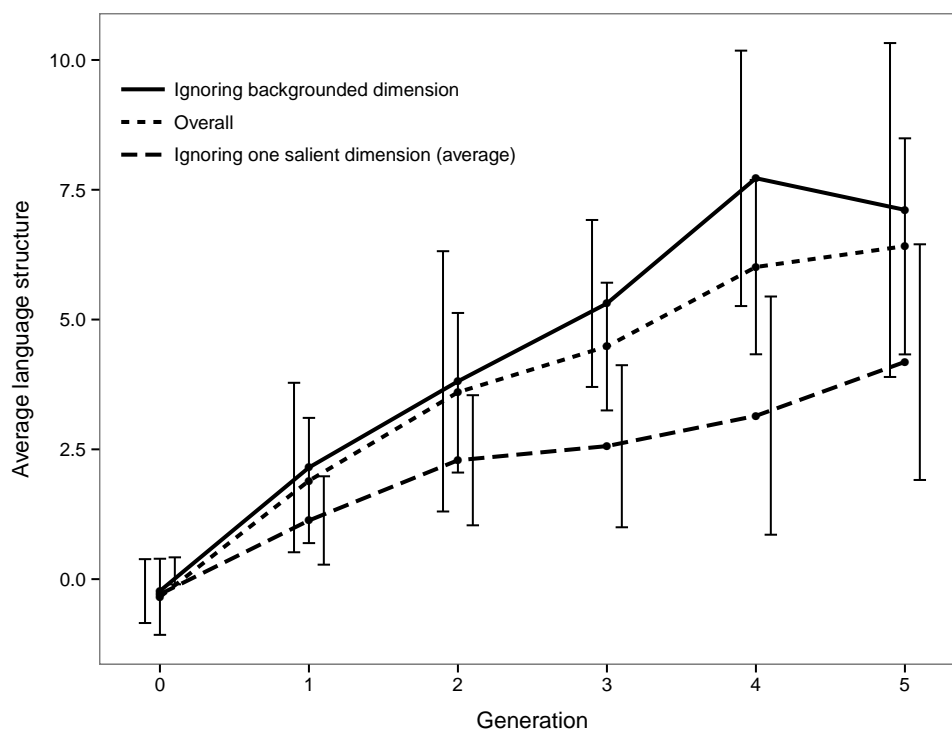


Figure 4.10: Change in structure (z -scores; see section 4.3.4.3 for how this measure is calculated) of the languages over generations. The dotted line shows structure calculated with respect to every meaning dimension. The solid line shows structure calculated with respect to the salient dimensions only, i.e., with the backgrounded dimension ignored. The dashed line shows structure calculated with respect to the backgrounded dimension and each of the salient dimensions, averaged. Error bars show 95% confidence intervals.

dimension versus pairs that differed only on a salient dimension are shown in Figure 4.9. Likelihood of collapse for backgrounded dimensions was 0.73, 95% CI [0.66, 0.80], whereas likelihood of collapse for salient dimensions was 0.38 [0.33, 0.43]. This was a difference of 0.35 [0.27, 0.43].

I analysed this with a mixed-effects logistic regression model, with Chain as a random slope over backgrounded/salient dimensions (as before, the inclusion of this random effect was assessed by likelihood ratios). The model found a significant effect of dimension type (backgrounded vs. salient) on the likelihood of collapse: $\beta = -1.52$, $SE = 0.28$, $z = -5.37$, $p < .001$. d_{unb} was 0.70, suggesting a large effect.

4.4.3 Language structure

Figure 4.10 shows the change in average language structure over generations. At generation 1, structure scores were broadly comparable regardless of which dimensions were taken into account: if all dimensions were taken into account, average structure score was 1.9, 95% CI [0.69, 3.10]; if the backgrounded dimension was ignored, average structure score was 2.15 [0.52, 3.78], a difference of 0.25 [-0.63, 1.13]. If one salient dimension was ignored, average structure score was 1.13 [0.28, 1.98]. This was 0.77 [0.11, 1.43] less than when all dimensions were taken into account, and 1.02 [-0.42, 2.46] less than when the backgrounded dimension was ignored. Thus, at this stage, ignoring the salient dimension has a medium negative effect on structure compared to a score that takes all dimensions into account; however, ignoring the backgrounded dimension does not result in a substantially different structure score.

These structure scores diverged over generations. By generation 4, if all dimensions were taken into account, average structure was 6.01 [4.33, 7.69]; if the backgrounded dimension was ignored, structure was 7.72 [5.26, 10.18]. This was higher by 1.71 [0.91, 2.51]. If one salient dimension was ignored, average structure score was 3.15 [0.86, 5.44]. This was a 2.87 [1.90, 3.84] lower than when all dimensions were taken into account, and 4.58 [2.83, 6.33] lower than when the backgrounded dimension was ignored. At this point, ignoring the backgrounded dimension results in the highest structure score; structure scores fall if it is taken into account, and fall further if one salient dimension is ignored.

A linear mixed-effects model examined how language structure was affected by generation and by which, if any, meaning dimensions were ignored when measuring structure. The mixed-effects model incorporated a random intercept for Chain, and fixed effects of Generation and dimension ignored (backgrounded, salient, or none). The model found that structure increased significantly over generations, $\beta = 5.62$, $SE = 0.85$, $t = 6.61$, $p < .001$. When one salient dimension was ignored, structure was significantly lower than when all dimensions were taken into account, $\beta = -1.50$, $SE = 0.43$, $t = -3.53$, $p < .001$. However, ignoring the backgrounded dimension did not significantly affect structure scores, $t = 1.30$, $p = .20$. The effect of Generation was significantly different for structure scores that ignored a salient dimension, as compared to structure scores that took all dimensions into account, $\beta = -2.19$, $SE = 1.04$, $t = -2.11$, $p = .04$. However, this interaction was not significant when comparing structure scores that ignored the backgrounded dimension and overall structure scores,

$t = 0.78$, $p = .43$, suggesting that the structure trend over generations was similar whether the backgrounded dimension was ignored or taken into account.

4.5 Discussion

4.5.1 Underspecification on backgrounded dimensions

As predicted, patterns of underspecification that reflected the salience of dimensions in learning and production contexts arose gradually over generations of cultural transmission. Starting from input languages that specified equally across all dimensions, the languages lost distinctions earlier and faster on the dimension that was consistently backgrounded during learning and use. Figure 4.11 shows a generation 5 language that underspecified more consistently on the backgrounded dimension (here, motion) than on the salient dimensions. This was typical of the final languages in the experiment.

The gradualness of the effect is a product of individual-level learning processes amplified by cultural transmission. The first participant in each chain learns a language that sends a strong signal that all distinctions on all dimensions are important (since each image is labeled by a unique word). The performance of these participants on the dimension selection task shows that they have absorbed this expectation: they select all dimensions equivalently, showing that they consider them equally important to word meaning. However, this 27-word language is not learnable within the constraints of the training regime. Therefore, when these participants have to reproduce the language in the test phase, they are frequently faced with situations where they do not recall the word for the target referent. In this situation, a sensible strategy is to reuse a word they remember to be associated with at least one of the features of the referent. The question is, which feature(s) will they choose?

Globally, the initial language treats all dimensions as equally important. However, when participants are actually learning the meaning of each word, attending to the backgrounded dimension will never improve their success in the discrimination game. This systematic manipulation means that the learner will tend to associate words more reliably with their referents' features on the more salient control dimensions than on the less salient backgrounded dimension. The analogous systematic structure of the production task, where the participant is cued to produce a word that will successfully discriminate the target from the distractor, also influences them to use a word which they associate with a feature on the salient dimension(s), rather than on the

backgrounded dimension.

Therefore, participants will tend to reuse words for referents that differ on the backgrounded dimension. The participant's task is still to converge on the language they are trained on, so errors in this direction will tend to be small and not necessarily systematic. However, as these errors build up over generations, they change the language and hence introduce a new source of evidence for the unimportance of one dimension: the patterns of word use in the language itself. Once a learner observes that a word can generalise over features on a dimension, this encourages the learner to reuse it for other cases if their memory fails. As more and more words underspecify on a particular dimension, learners become more willing to extend this pattern of underspecification to other words in the language (see Figure 4.11 for a generation-by-generation view of how underspecification on the backgrounded dimension spreads as a chain progresses).

The pattern of results from the dimension selection task is also illuminating. Even though the language participants were tested on in this task was the language they were trained on, not the language they produced, participants' behaviour better reflected the patterns of underspecification in the language they produced. If participants were instead reflecting their training language, we would expect a smaller difference in preference of salient over backgrounded dimensions in generation 3 (where the generation 2 input language had only moderate levels of underspecification on the backgrounded dimension), and a larger difference in generation 5 (where the generation 4 input language had the highest levels of underspecification on the backgrounded dimension in the whole experiment). The fact that these results so closely reflect the language-based measures of underspecification both validates the results and shows that participants' expectations about the relevance of particular dimensions changed during their production of the language, relative to the encoding of the dimensions in the training language. This provides further confirmation that the change that happens during recall is not just undirected error, but evidence of change in the participant's representation of the word meanings in the language.

4.5.2 Partial compositionality

One output language showed an interesting variation on the typical pattern (Figure 4.12). While it still became underspecified on the backgrounded dimension, the other two dimensions were encoded in a compositional system, with the meaning of each label a function of the meanings of the label's parts. For example, the substring 'ti-'/'ta-

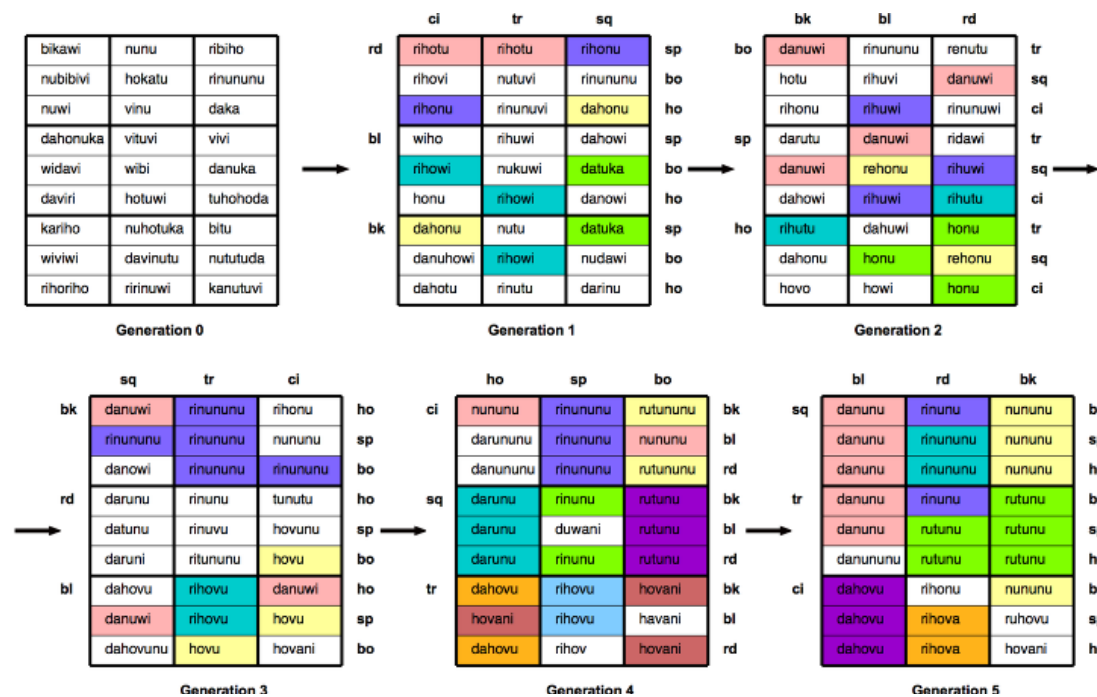


Figure 4.11: The emergence of underspecification in chain 7. Each grid shows one participant’s language, arranged so that the backgrounded dimension always runs down the right-hand side. Abbreviations as in legend to Figure 4.2. Words used for more than one referent, i.e. underspecified words, are filled with the same colour. The thick gridlines indicate regions that would be filled with the same colour if the language underspecified on the backgrounded dimension. The figure shows underspecification arising more consistently on the backgrounded dimension than on either of the salient dimensions, although it also extends partially to the salient dimensions (see e.g. the ‘overspill’ of pink and green regions in generation 5).

	sq	ci	tr	
bl	ta-ho	ti-kiwiki	ta-kiwi	bo
	ta-ho	ta-wiki	ta-ho	ho
	ta-ho	ti-kiwiki	ta-kiwi	sp
rd	bo-ho	bo-kiwiki	ru-kiwo	bo
	bo-ho	bo-wiki	ru-biwo	ho
	bo-ho	bo-wiki	ru-kiwo	sp
bk	da-ho	da-kiwiki	da-kiwi	bo
	ho-ho	da-wiki	da-kiwi	ho
	ho-ho	da-wiki	da-kiwi	sp

Figure 4.12: A partially compositional language that emerged in generation 5 of chain 3. As in Figure 4.11, cells filled with the same colour correspond to images referred to by the same word. Substrings are separated with hyphens for clarity; as typed in the experiment, words had no hyphens.

corresponds to blue, and the substring ‘-ho’ corresponds to square, creating ‘taho’ meaning blue square. While this system is not entirely consistent (see, for example, the variation between ‘bo-’ and ‘ru-’ as substrings for red, and ‘-wiki’ and ‘-kiwiki’ as substrings for circle), it is completely inconsistent on the backgrounded dimension, where the few distinctions are unsystematic (e.g., ‘rubiwo’ for red triangle moving horizontally, versus ‘rukiwo’ for the other possible motions, with this b/k alternation not occurring in any other context). This kind of compositional system, an alternative solution for increasing learnability while maintaining expressivity, was also observed in Kirby et al. (2008)’s Experiment 2. The intriguing outcome in the current experiment is that even when an adaptive solution like compositionality emerges, it does not encode the backgrounded dimension. This underlines the difference between an underspecification that emerges purely from memory constraints and a lack of pressure to communicate unambiguously, and a motivated underspecification cued by systematic features of contexts of learning and use.

The emergence of this partial compositionality also shows that the inclusion of a given feature as part of a word meaning, and therefore the likelihood of its being encoded as part of a language-wide compositional system, depends on its salience in situations of learning and use. This is an important finding for investigations of the emergence of linguistic structure that assume a structured meaning space: if a given perceptible dimension of the meanings is habitually less salient in situations of learning and use, an underspecified language may be a more efficient solution than a fully compositional one that encodes the less salient features.

4.5.3 Language structure

As shown in Figure 4.10, structure in the languages increased over generations, replicating Kirby et al. (2008). However, this structure did not significantly drop when the backgrounded dimension was ignored. This finding supports the general result that habitual backgrounding of a dimension in situations of learning and use makes it less likely, over generations of cultural transmission, that language users will condition their patterns of word reuse on this dimension.

4.5.4 Problems and future directions

4.5.4.1 Underspecification on salient dimensions

The majority of the output languages in generation 5 were underspecified across more than just the backgrounded dimension. The generation 5 language in Figure 4.11 underspecifies not only across the motion dimension (the backgrounded dimension for this participant) but also partially across the shape dimension – e.g., a blue square and a blue triangle are both called ‘danunu’. This shows that the undirected underspecification that arose in Kirby et al. (2008) also occurred in this experiment, in addition to the underspecification cued by the experimental manipulation. The learning- and testing-based procedure in this experiment, while sufficient to direct underspecification preferentially toward backgrounded dimensions, is not sufficient to prevent it eventually arising on salient dimensions. However, the collapse of distinctions on salient dimensions is still significantly lower overall than on backgrounded dimensions, as shown by the diachronic results in Figures 4.8 and 4.9.

In real language, other pressures presumably prevent undirected underspecification from occurring: for example, a pressure for unambiguous communication. One avenue for future work is to explore whether, with the introduction of a pressure for unambiguous communication (following Kirby et al., submitted), underspecification would still emerge on the backgrounded dimension, while distinctions on salient dimensions would be preserved. Experiments 3, 4 and 5, presented in Chapters 5 and 6 of this thesis, investigate the extent to which a pressure for communication encourages meaning distinctions to be maintained.

4.5.4.2 Generation 5

As seen in Figures 4.5, 4.6, 4.8, 4.9 and 4.10, average underspecification on backgrounded dimensions actually dropped in generation 5, going against the overall trend and the hypothesis. The data were analysed chain by chain to find the cause of this decrease.

3 out of 8 chains (chains 1, 6, and 7) were consistent with the hypothesis, losing distinctions on the backgrounded dimension from generation 4 to generation 5. The remaining chains gained distinctions on the backgrounded dimension to varying degrees. The language in chain 3 (the compositional system shown in Figure 4.12) gained 3 distinctions on the backgrounded dimension and lost 2, a net increase of 1 distinction.

Error was low (0.22) and the added distinctions were a result of inconsistent use of the -wiki/-kiwiki suffix for circle. The language in chain 8 gained 5 distinctions and lost 2, a net increase of 3. However, some of these distinctions looked to be the result of typos in otherwise consistently used words: e.g. ‘op’ for ‘opa’, ‘kanuwuru’ for ‘kaniwuru’. The languages in chains 4 and 5 each lost 1 distinction on the backgrounded dimension and gained 5, giving a net increase of 4. Error in these chains was average (0.35, 0.24), and there is nothing strikingly unusual about the changes made by the learners. Finally, the language in chain 2 gained 5 distinctions on the backgrounded dimensions. However, this chain also had the highest error in generation 5 (0.64, 2 standard deviations above the mean of 0.35 and closer to the average value for generation 2, 0.63). In addition, this participant stated after the experiment that they had made up their own systematic language rather than trying to learn the one presented.

In conclusion, the small number of chains makes it hard to be certain whether the drop in underspecification on the backgrounded dimension in generation 5 is a systematic effect or a quirk of particular participants. For chains 4, 5 and 8 in particular, the small changes and low error are typical of iterated learning, so the fact that these chains go against the hypothesis should not be dismissed out of hand: it is possible that the underspecification on backgrounded dimensions that emerges over the course of the experiment is not stable. The best solution would be either to run further chains and see if a larger sample produces generation 5 results that fit better with the overall trend, or to run the chains for longer and see if underspecification on the backgrounded dimension stabilises.

4.5.4.3 Validity of experimental manipulation

As outlined in section 4.2 above, the experimental manipulation, of one dimension never being useful for the discrimination game, was intended as a proxy for other factors that might make particular dimensions more or less salient in context of language learning and use. However, the potential criticism applies that this manipulation does not have an obvious real-world counterpart. Experiment 5, presented in Chapter 6, manipulates the salience of different dimensions in a potentially more ecologically valid way, by presenting possible events in an artificial world where features on different dimensions co-occur more or less predictably. Picking up on the lack of division between the lexicon and world knowledge discussed in Chapter 2, this experiment investigates how possible events in the world form an ‘implicit context’, cueing some dimensions over others as more useful to specify.

4.5.4.4 Known salience effects of particular dimensions

As the discussion of the shape bias in the introduction to this chapter makes clear, different stimulus dimensions have different prior salience: participants' expectation that noun labels are more likely to generalise by shape than by colour likely affects their behaviour in the experiment. The shuffling process described in section 4.3.3.2 was intended to compensate for these effects; however, this process merely erases these effects at every generation, preventing them from building up into a cumulative effect on the language, rather than preventing them from occurring at all. It is likely that the specific dimension being backgrounded affected how strongly underspecification emerged for each participant. For example, motion was the backgrounded dimension in 5 out of 8 chains in generation 5, which might have contributed to the odd results in this generation. To check this, the analyses reported in the results were re-run including which dimension was backgrounded as a fixed effect; however, in no case did this improve the fit of the model, and within these models the specific dimension being backgrounded was not a significant predictor of underspecification. This is reassuring, as it suggests the differing prior salience of different dimensions did not have a systematic effect on the results. However, to enable easier differentiation between dimension-specific effects and the general effect of the experimental manipulation, it would have been better to counterbalance, rather than randomise, which dimension was the backgrounded dimension, so that each was backgrounded in an equal number of chains at each generation.

In part to address these concerns, Experiments 2, 3 and 4, presented in Chapter 5, move away from a dimensional feature-based model of meaning. These experiments instead use a continuous Euclidean space of morphed images, allowing investigation of how word meanings arise and develop where perceptual structure is more unclear or ambiguous.

4.6 Conclusion

In this experiment, I set out to investigate the origins of patterns of underspecification that help learners generalise words appropriately. The results show that attentional learning effects amplified over cultural transmission lead to a lexicon that underspecifies preferentially across dimensions that are habitually less salient during learning and use. Thinking of the language in the experiment as analogous to a particular region of

the lexicon, for example object or substance nouns, illuminates a possible mechanism for the origin of the strong tendencies to specify on particular dimensions we see in these regions. Over a whole language, specification will range over various dimensions depending on the functions of individual words, as well as the characteristic situations in which they are learned and used. For example, the relational nature of gradable adjectives such as ‘big’ means that the contexts in which they are learned and used will tend to highlight dimensions of relations between objects as well as intrinsic dimensions (Clark & Amaral, 2010; Gentner & Kurtz, 2005; Sandhofer & Smith, 2001). More broadly, a language can be seen as a dynamic system where the meanings of individual words adapt to, as well as themselves contributing to, the salience of particular dimensions in contexts of learning and use. This result uncovers a mechanism for how words can come to specify in adaptive ways: as a cumulative product of the incremental changes made by individual learners attending to contextual cues in learning and use.

More broadly, the results show how attentional learning and iterated learning interact to shape patterns of word reuse. In this experiment, these two pressures work complementarily: the bottleneck imposed by iterated learning, i.e., the fact that participants cannot remember and reproduce all the word-referent pairings they are exposed to, leads to situations where participants must reuse known words for novel referents. Attentional learning then pushes participants to condition their word reuse on a) dimensions that are more salient in the context of use, and b) dimensions that condition the patterns of reuse of other words in the language. The results of this experiment thus demonstrate how adaptive regularities in the patterns of reuse of words across a lexicon could have emerged without having to stem from the matching of words to pre-linguistic concepts. A lexicon that assists learners in making higher-order generalisations about the appropriate dimensions on which distinctions should be made can itself arise as a product of cultural evolution. Individual learning effects built up over generations create patterns that help new learners to make the appropriate generalisations about the dimensions on which they should condition their reuse of words.

Learning, communication, and the structure of word meanings: Convexity and alignment increase over iterated learning and communication¹

5.1 Introduction

Categorisation is the process of making a continuous world discrete. Rather than treating all stimuli as unique, we recognise them as members of previously encountered groups. This is an adaptive way of generalising learned experiences: if you recognise a new stimulus as a member of a familiar category, you can make useful inferences about how to respond (Markman & Ross, 2003). As outlined in Chapter 2, categorisation is advantageous even before language is taken into account: ‘With or without language, the mind has to have a way to unify multimodal representations and store them as units’ (Jackendoff, 2002, p. 349). Animals ranging from pigeons (Watanabe et al., 1995) to baboons (Fagot et al., 2001; Bergman et al., 2003) engage in categorisation behaviour, showing that this phenomenon is not dependent on language: it can result from individual learning biases applied to a structured world, where phenomena

¹Part of this chapter is published as Silvey et al. (2013).

are grouped according to important (e.g., survival-relevant, Noh et al., 2014) dimensions.

Word meanings are another way of making the continuous discrete. By reusing a word to apply to a novel referent, we increase the range of what we can communicate about without having to learn an infinite number of words. The benefit of this behaviour seems intuitively similar to the benefit of categorisation for individual learning described above. Indeed, as outlined in Chapters 2 and 3, mainstream views have equated word meanings with the concepts that define individually learned categories, arguing that words merely label these pre-existing groupings. However, I argue that these two kinds of discretisation – individual categorisation and communicative word reuse – come about via radically different mechanisms. As such, we should expect patterns of communicative word reuse to be subject to different pressures from those acting on individually learned categories. As discussed in section 3.5.3, we might expect communication to give rise to patterns of word reuse that enable finer, more expressive distinctions (Pinker & Bloom, 1990; Kirby et al., submitted); that patterns of word reuse will be structured optimally for communication (Freyd, 1983; Gärdenfors, 2000); and that pairs who communicate together will end up with patterns of word reuse that align (Steels & Belpaeme, 2005; Garrod & Pickering, 2009; Voiklis & Corter, 2012). Furthermore, unlike individually learned categories, word meanings are not induced directly from the world at each new generation, but are learned and passed on in a process of cultural transmission. As outlined in section 3.5.2 of Chapter 3, this process of iterated learning may work against communicative pressures to reduce the number of meaning distinctions that can be made, and to shape the patterns of reuse of words to be more easily learnable.

This chapter presents a series of experiments that test these hypotheses. The questions addressed are: Does communication lead to word meanings that are differently structured from individually produced categories? Does iterated learning interact with communication to produce word meanings that are more learnable, and if so, are these learnable meanings also those which are most useful for communication? In the experiments presented in this chapter, participants divide a novel, continuous space of stimuli into labelled categories which serve as a proxy for word meanings. In Experiment 2, participants rate the pairwise similarity of images in the stimulus space to establish a baseline category system. Experiment 3 compares category systems produced by individuals with category systems produced by communicating pairs. In one condition, pairs of participants divide the stimulus space into categories via their patterns of word

reuse over the course of a communication game; in the other two conditions, participants create categories individually, with these two conditions varying on whether categorisation decisions are made simultaneously or sequentially. While communication does lead to more expressive systems (i.e., systems with more labelled categories), surprisingly, the category systems produced during communication are less aligned, and structured in a way that is less optimal for communication, than systems produced by individuals. In Experiment 4, communicating pairs first learn a category system, then use it in a communication game, and finally pass this system on to be learned by subsequent pairs in an iterated learning chain (as in Experiment 1). This combination of repeated communication and learning leads to category systems that increase in structure and alignment over generations. While the expressivity of the systems drop in response to learning pressures, communicative systems remain consistently more expressive than systems produced by individual categorisation.

5.2 General Method

This section describes the stimuli used in the experiments, and the dependent variables used to assess the structure of participants' category systems. It also details Experiment 2, in which participants' pairwise ratings of stimulus similarity were collected and used to produce a baseline similarity-based categorisation of the stimulus space.

5.2.1 Stimuli

Experiment 1 (Chapter 4) used stimuli with feature-based structure on perceptually obvious dimensions. While these stimuli have some scope for flexible construal, in that distinctions on particular dimensions can be marked or ignored, the available distinctions are discrete: for example, each shape is either a square, a circle or a triangle, with no values in between. As discussed in section 3.5.1 of Chapter 3, learning and communication interact with world structure to influence the structure of word meanings. By adopting a meaning space whose structure is less discrete, we can examine the effects of learning and communication without strong pre-existing world structure determining the word meanings that evolve. If the appropriate placement of category boundaries in the stimulus set is ambiguous, this leaves room for variation in how participants divide this continuous space into discrete sets. Use of such 'alternative materials' also forces participants to create features for categorisation, rather than have these features

be obviously available a priori (Schyns et al., 1998).

The stimulus space used is therefore a set of morphed images, generated by the same method as those in Matthews (2009)'s iterated categorisation experiment cited in Chapter 3. The set of images is shown in Figure 5.1. In pilot studies, participants showed variation in where they drew category boundaries between these images, making them suitable for the experiments. The four corner images (images 0, 4, 20 and 24) were generated using PsychoPy (Peirce, 2007), via the following method: For each corner image, a random number generator assigned x and y positions for the five vertices, and the resulting shape was drawn by connecting these vertices. Morphs between these images were then generated to fill the rest of the space. For a given image x in the set, the position of each vertex was an average of the positions of the corresponding vertices of the corner images, weighted according to x 's inverse Euclidean distance from each corner. This resulted in a total set of 25 images whose similarity and difference are defined by Euclidean distance. Peter Gärdenfors argues that 'conceptual spaces can be modelled as Euclidean spaces' (Warglien & Gärdenfors, 2011, p. 2170), and his claim that natural concepts are convex regions can be easily tested in this kind of space (see section 5.2.2 below). In addition, the Euclidean space provides a basis for success scores in the communication game: during communication, interlocutors are rewarded the more similar the image they pick is to the image their partner intended.

The objective Euclidean distance between the images in the space may of course not correspond directly to perceptual similarity (see, e.g., L. Smith & Heise, 1992). To check the extent to which similarity perception reflects the Euclidean space, Experiment 2 was run to collect pairwise judgements of stimulus similarity of images in the set. From these ratings a baseline category system was constructed to serve as a reference for the experimental analysis.

5.2.2 Outcomes

To test the hypotheses outlined in the Introduction, we need ways of measuring the expressivity, structural optimality, and alignment of the category systems produced in the experiments.

5.2.2.1 Expressivity

A useful proxy for the expressivity of a system is the number of categories it contains, i.e., the number of distinct regions into which the system divides up the continuous

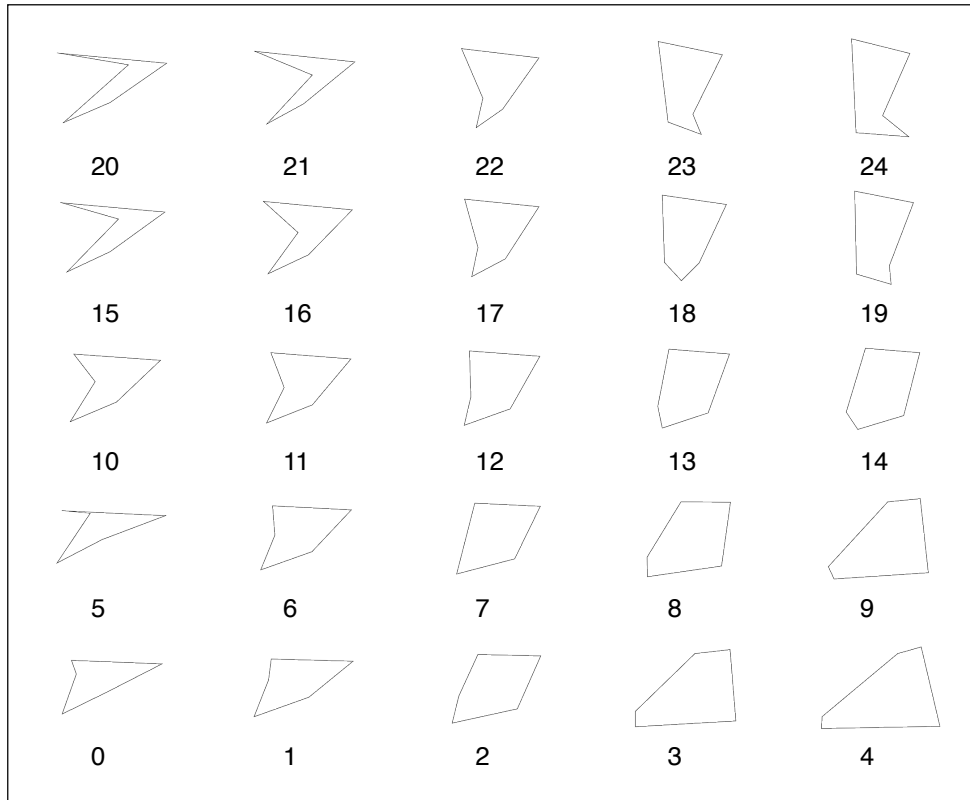


Figure 5.1: Continuous space of morphed images used as stimuli in Experiments 2, 3 and 4. For the purposes of the communication game in the experiments, success scores are proportional to inverse Euclidean distance. For example, picking image 0 when the target is image 0 gains 15 points (maximum); picking image 24 when the target is image 0 gains 1 point (minimum).

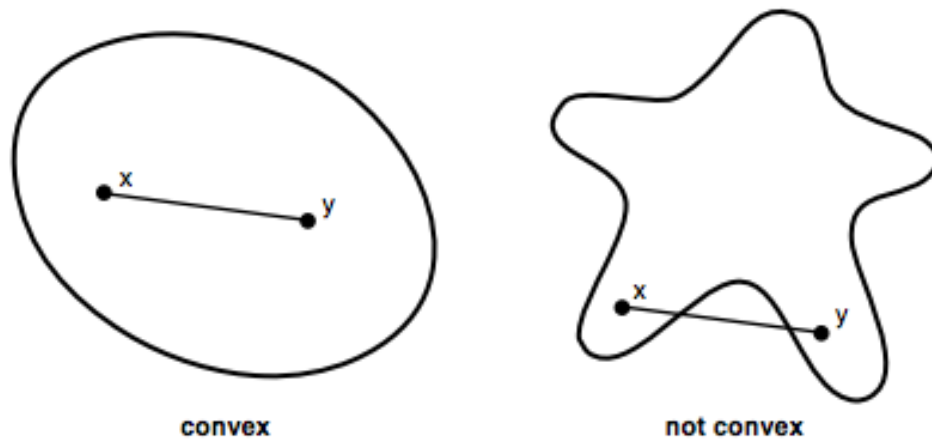


Figure 5.2: Illustration of convex and non-convex regions of conceptual space, taken from Gärdenfors (2000).

space. Number of categories is used to measure expressivity in all analyses.

5.2.2.2 Category structure: convexity

As discussed in Chapter 3, some theorists have argued that particular category structures may be more cognitively natural and/or more useful for communication. In particular, I here investigate the hypothesis of Peter Gärdenfors that word meanings correspond to natural concepts, which are defined as being convex regions of conceptual space (see Figure 5.2 for an illustration). Convexity is defined as follows:

A subset C of a conceptual space S is said to be *convex* if, for all points x and y in C , all points between x and y are also in C . (Gärdenfors, 2000, p. 69).

Gärdenfors's proposal is motivated by 'a principle of *cognitive economy*; handling convex sets puts less strain on learning, on your memory, and on your processing capacities than working with arbitrarily shaped regions' (Gärdenfors, 2000, p. 7). The reason these regions are economical is as follows: given a Euclidean conceptual space, a number of prototypes, and a rule that novel exemplars should belong to the category defined by the closest prototype in the space, the categorisation generated is a Voronoi tessellation, where each category is convex by definition (Gärdenfors, 2000). In this way, convex categories can be generated on the basis of prototypes and the similarity gradient of the conceptual space, rather than every category member having to be learned individually. Gärdenfors cites evidence of word meanings in real languages that are convex, for example colour categories in Swedish (Sivik & Taft, 1994)

and spatial prepositions in a number of languages (Zwarts, 1995; Bowerman & Choi, 2001).

Gärdenfors's account focuses on the cognitive motivation for convex concepts, in that they allow efficient generalisation from limited exemplars. However, he also entertains the possibility that convex concepts may arise specifically from communication, via the mechanisms by which we generalise words to novel situations.² Gärdenfors notes the 'chicken and egg problem' of whether categories with this kind of structure are a prerequisite for, or a product of, successful communication (Gärdenfors, 2000, p. 196). Evidence supporting communication as a potential source for convex category structures comes from the work of Gerhard Jäger, who finds that in a simulated communication game with a Euclidean meaning space, senders and receivers converge on convex regions of the space as meanings (Jäger, 2008). Jäger argues on this basis that 'the convexity of cognitive categories is not so much a property of cognition but rather a consequence of a positive feedback loop in communication' (Jäger, 2008, pp. 562-563). However, he acknowledges that the parameters of the simulation also allow a non-communicative interpretation, whereby 'the convexity of categories is simply the consequence of a tendency to maximize the similarity of the instances of the same category' (p.563). Thus, there is some disagreement about whether convex categories arise from properties of individual cognition or from mechanisms of communicative interaction. The studies presented in this chapter investigate this problem empirically.

It should be noted that the convexity hypothesis comes with a number of strong assumptions. The first is a conceptual space that is geometrically structured: specifically, a space where dissimilarity between stimuli can be modelled as Euclidean distance, making it meaningful to characterise a given point as being 'between' another pair of points in the space. Gärdenfors himself notes that this assumption may not hold for all real-world conceptual spaces. The idea of a word meaning as a region of conceptual space is also at odds with the context-dependent, exemplar-based model of meaning argued for in Chapter 2. Gärdenfors himself acknowledges the context-dependence of concepts: 'what appears to be a core property in one context may seem peripheral in another' (Gärdenfors, 2000, p. 107). However, he argues that his model can accommodate this context-dependence by varying the weights assigned to differ-

²Gärdenfors cites Jennifer Freyd's shareability hypothesis (discussed in Chapter 3) in reference to this idea; however, Freyd's hypothesis is that novel referents will be conceptualised in a way that makes them easily describable by existing conventions, even if this involves distorting their perceptual features. The shareability hypothesis therefore does not specifically predict that words will refer to convex regions of conceptual space.

ent conceptual domains. In Gärdenfors's account, then, word meanings are convex *in context*, i.e., in a contextually defined conceptual space, rather than one unchanging, monolithic space. In this way, Gärdenfors acknowledges the extensive work in the concepts and categories literature showing that similarity is not constant but is context-dependent (Tversky, 1977; Nosofsky, 1986), and can itself be affected by categorisation (Goldstone, 1994) and labelling patterns: 'any similarity metric must be capable of significant modulation by the distribution of labels over parts of a representational space' (Landau & Shipley, 2001, p. 109). However, experimental methods allow us to abstract away from this context-dependence, providing an opportunity to examine the effects of similarity as one of many contextually constrained contributors to word meaning structure. For example, the paper by Landau and colleagues cited above shows that, in a controlled experimental context, learners will tend to generalise words to convex regions of the stimulus space. Adapting the convexity hypothesis to the model of word meanings presented in Chapters 2 and 3, then, we can ask whether exemplars labelled by the same word will tend to cluster in convex regions of the stimulus space, and whether the convexity of these regions will vary depending on whether the labelling is done individually (suggesting a general cognitive mechanism) or in the context of communication (suggesting a specifically communicative mechanism).

Applying Gärdenfors's definition of convex categories to the stimuli in Figure 5.1, a convex category is one where it is possible to draw a line between the centres of any two members of the category and not cross the centre of any image in a different category. There are several possible ways to quantify this. One is simply to code whether each category in a system is convex (1) or non-convex (0), and then average these values to give the proportion of categories in a system which are convex. However, some non-convex categories may be closer to being convex than others: for example, having only one outlier, rather than being spread unevenly across the stimulus set. A binary analysis fails to take account of this variation. As an alternative, I use the proportion of the neighbours of each image i that are in the same category as i to define a measure of graded convexity. Rather than simply classifying each category as convex or not, this measure allows us to quantify the extent to which a given category (and, by averaging over categories, a whole system) approaches convexity.

A caveat to the analysis here is that it assumes that the Euclidean distance metric that was used to generate the stimuli corresponds to the relative similarity of the stimuli as perceived by participants. The validity of this assumption, as well as contextual effects on similarity, are discussed in section 5.6 below.

The index used is adapted from Theiler & Gisler (1997). Over all members of a category, the proportion of members' neighbors also in the category is averaged to provide a convexity measure for the category. The average of these category values is then taken as the convexity measure for the whole system. Categories with only one member are not included in this calculation, since they are neither meaningfully convex nor non-convex.³

A problem with this measure is that its possible values vary with the number of categories in a system and the number of images assigned to each category. To make the index comparable across different category systems, it was adjusted using the following general formula (Hubert & Arabie, 1985):

$$\frac{\text{Index} - \text{Expected Index}}{\text{Maximum Index} - \text{Expected Index}}$$

The Expected Index (i.e. the convexity of a random system with the number of categories and distribution of images across categories of the system being analysed) was obtained via simulation: for a given category system, we generated 900,000 random category systems by shuffling the assignment of images to categories, while keeping the number of categories and number of images per category constant. The convexity measures for these shuffled category systems were averaged to give the expected level of convexity for this system.

The Maximum Index (i.e. the maximum possible convexity index for a given system) was also obtained by simulation, in this case a search procedure implemented using a genetic algorithm. In order to find the maximally convex variant of a given category system, we ran an evolutionary simulation, evolving a population of 1000 shuffled versions of the veridical category system. At each generation of the simulation, the 500 category systems that produced the highest convexity scores were 'mutated' (swapping two of the image-to-category assignments at random, while keeping the number of categories and the number of images per category constant). These mutated category systems, along with their 'parents', became the new population. The simulation was run three times, for 1000 generations, at which point all 3 runs reliably produced an estimate of the maximum convexity index which agreed to within 2 decimal places of precision.

³A system with only one category containing all images would also be neither meaningfully convex nor non-convex; however, since none of the systems produced in the experiments had fewer than 3 categories, this was not an issue for the dataset.

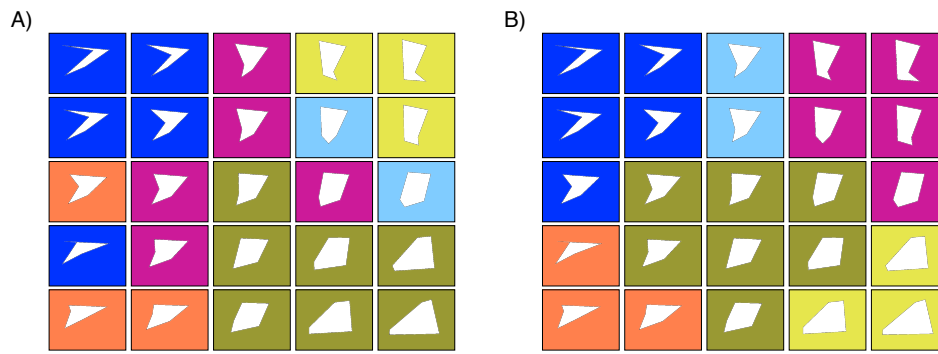


Figure 5.3: A) An example category system produced during Experiment 3. Images with the same background colour were placed in the same category. In the experiment, colours correspond to words; in this example, the assignment of colours to particular groups of images is non-meaningful. B) The same category system, with assignments of images to categories shuffled to produce the most convex arrangement, according to the measure specified above. The convexity index for this system was used as the ceiling to calculate the adjusted convexity measure. The value of the adjusted convexity measure for the system on the left is 0.65, while the value for the system on the right is 1.

Figure 5.3 shows an example of a veridical category system and a shuffle producing the maximum convexity index. From Figure 5.3, we can see that the shuffled system with the maximum possible convexity contains only categories that are convex, by Gärdenfors's definition. Therefore, while this measure does not start out by directly measuring convexity, maximum values of this measure do correspond to maximally convex categories, making it a useful proxy. In addition, the hypothesis that convex categories are more optimal for communication holds in this stimulus space: the more closely clustered a category's members are in the Euclidean space, the more participants are able to maximise their score for a given target.

The convexity values reported in the Results are the veridical indices corrected according to the formula above. The resulting convexity index can be interpreted as follows: a convexity of 0 indicates the categories are no more convex than would be expected by chance⁴; a convexity of 1 indicates that the categories are maximally convex, given the possible arrangements of that particular category system.

⁴The index can take on small negative values in cases where categories are less convex than chance would predict.

5.2.2.3 Alignment

Several methods exist in the literature for quantifying alignment between category systems. An intuitive and frequently used measure is the Rand index (Rand, 1971). This index calculates the proportion of all possible pairs of images which participants agree should be placed either in the same or in different categories. Like the convexity measure discussed above, the Rand index's possible values are constrained by the characteristics of particular category systems (for example, the measure will tend to 1 as the number of categories increases). Therefore, the Adjusted Rand index (Hubert & Arabie, 1985), calculated by the same formula as above, was used. Details of how the chance value is calculated can be found in Hubert & Arabie (1985). The resulting measure is bounded⁵ from 0 to 1, where 0 = category systems are no more aligned than would be expected by chance, and 1 = perfect alignment, i.e. the category systems are identical.

An information-based metric was also considered: AMI_{\max} , proposed by Vinh et al. (2010). This measure is based on the mutual information between two category systems, normalised by the maximum of the entropies of the two systems so that it ranges from 0 to 1, and corrected for chance along the same lines as the Adjusted Rand Index (i.e., by calculating the expected value of mutual information for two systems with given numbers of categories and distribution of images per category, and adjusting according to the formula quoted in section 5.2.2.2). This metric proved to have values almost identical to those of the Adjusted Rand Index. Since it is the more frequently used measure, the Adjusted Rand index was adopted as the measure of alignment.

Since the variable of interest was participants' category systems rather than the words they used, the alignment measure was taken irrespective of label: i.e., if two participants put the same group of images in a category but used different labels, they would count as fully aligned for this category.

5.2.3 Experiment 2

To provide a baseline for assessing the category systems produced by participants in Experiments 3 and 4, it is desirable to obtain a categorisation of the images on the basis of similarity, without participants being explicitly instructed to categorise and without them using labels. To obtain this baseline categorisation, *k*-means clustering

⁵As with the convexity measure, the lower bound is stochastic; in practice, the measure can take small negative values if veridical alignment is less than would be expected by chance.

(Lloyd, 1982) was used to construct a category system on the basis of pairwise similarity judgements of the stimuli, obtained from participants in an online experiment.

5.2.3.1 Method

Selection of image pairs for comparison With 25 images, the total number of pairs is $n(n - 1)/2 = 300$. This is an impractically large number for a short online experiment. A number of solutions have been proposed for the problem of collecting a sufficient number of pairwise similarity judgements. The solution used here is an incomplete cyclic design (David, 1963). This design uses a subset of all possible pairs, selected under the following constraints: 1) each image must appear in the same number of pairs, and 2) each image is compared with each other image either directly or through a chain of comparisons (i.e., there is no way of grouping the images such that no member of one group is ever compared to a member of the other group).

The subset of pairs is selected by shuffling the list of stimuli, placing them in a circle, and then selecting sets of pairs based on their relative placement around the circle. Figure 5.4 shows an example. The more iterations of this process are performed, the higher the value of r , where $r =$ the number of paired comparisons each image appears in, and the higher the percentage of the total data is used. For Experiment 2, this process was performed 5 times, yielding $r=10$ and 125 pairs (just over 40% of the 300 possible pairs). Burton (2003) showed that results from designs using 40% of the data had mean .98 correlation with results from designs using complete data, justifying this choice.

Participants Participants were unpaid volunteers, recruited via Twitter and Facebook. 73 participants began the experiment; of these, 27 did not finish and so were excluded. Of the remaining 46 participants who provided complete data, the 6 oldest were excluded in order to keep the median age close to that of participants in Experiments 3 and 4. The final analysis group of 37 participants had a median age of 26. 28 were female and 1 preferred not to disclose their gender.

Procedure The experiment took place online, via an HTML/PHP interface. The general procedure was a standard one for collecting stimulus similarity judgements (Ekman, 1954; Abelson & Sermat, 1962). On each trial, participants were presented with a pair of images, with their placement to left and right randomised. Participants were asked to rate the similarity of the two images on a discrete scale from 1 (least

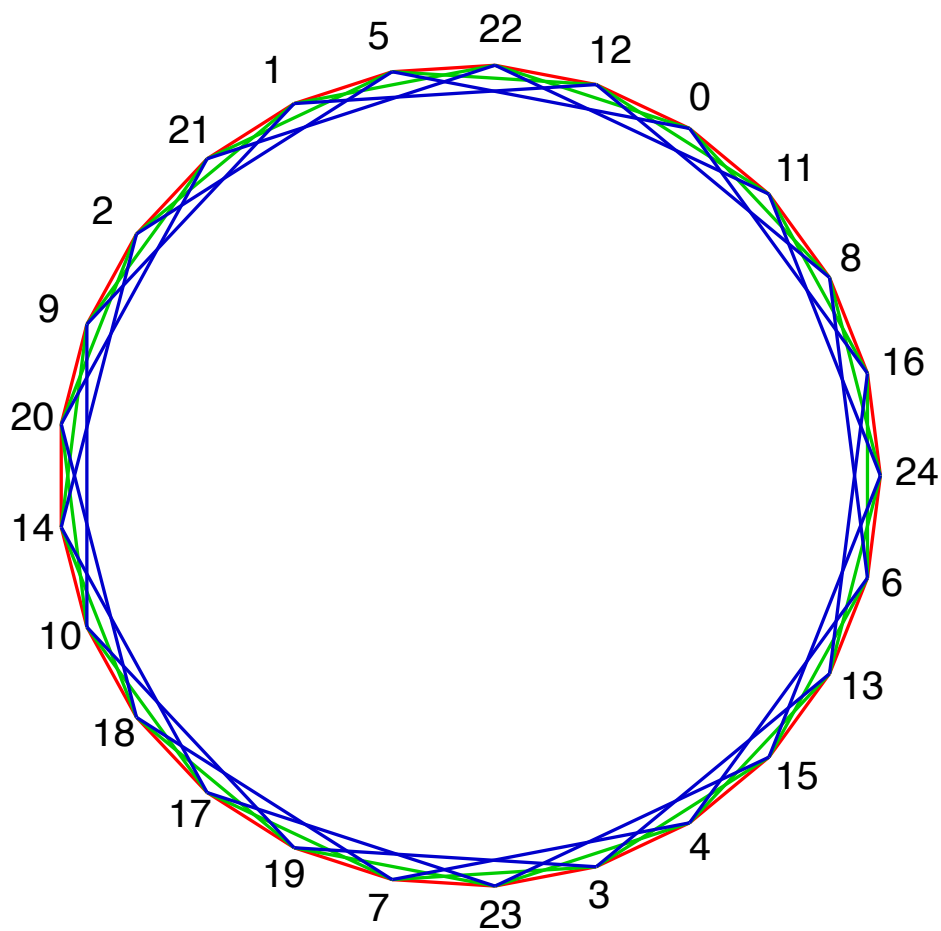


Figure 5.4: A method of selecting pairs for an incomplete cyclic design. Numbers refer to images in the stimulus set shown in Figure 5.1 above. The image numbers are shuffled and placed in a circle. The first set of pairs consists of each adjacent pair in the circle (red lines): for example, from the top clockwise, 22 is paired with 12, 12 is paired with 0, etc. The second set of pairs consists of each alternate pair in the circle (green lines): for example, 22 is paired with 0, 0 is paired with 8, etc. The third set of pairs consists of every third image (blue lines): 22 is paired with 11, 11 is paired with 24, etc. All of the illustrated sets of pairs are cyclic, i.e., there is no way of grouping the images such that no member of one group is ever compared with a member of the other group. Note that for this dataset, where $n=25$, the set of pairs produced by pairing every fifth image would not be cyclic. For clarity, this figure illustrates $r = 6$, i.e., each image appears in 6 paired comparisons; however, the design for Experiment 2 used $r=10$, meaning two further sets of pairs were included by drawing more lines connecting pairs further apart on the circle. This results in a set that uses over 40% of all possible pairs.

similar) to 7 (most similar). There were 125 trials in total. The experiment took around 10 minutes to complete.

5.2.3.2 Results

Similarity ratings were converted to dissimilarity ratings for the purpose of *k*-means clustering: i.e., if a pair was rated 7 by a participant (= most similar), this score was converted to a 1, and vice versa proportionally along the scale. These dissimilarity scores were then averaged over participants for each pair of images tested. The resulting incomplete dissimilarity matrix contained an average dissimilarity score for each pair of images that appeared in the experiment, plus missing values for those pairs that did not appear. These missing values were imputed as the average of the remaining data points, as in Burton (2003).

This matrix was then analysed using *k*-means clustering. This algorithm takes dissimilarity ratings as input and uses an iterative procedure to find the category system that most closely satisfies the following properties: a) each category has a centre which is the prototype for that category; b) each item is assigned to the category whose centre it is closest to. This was implemented via the *k*-means algorithm from the stats library in R (R Core Team, 2013). The appropriate number of clusters was first assessed by plotting the within groups sum of squares for each number of clusters and using the ‘elbow method’ to find the smallest number of clusters where adding another cluster did not result in significant improvement in fit (Everitt & Hothorn, 2010). This analysis was run multiple times to check the consistency of the ‘elbow’: while it sometimes appeared at 5 or 7, overall it was most predominant at 6, suggesting that dividing the images into 6 clusters maximised improvement in fit. Therefore the *k*-means algorithm was run to find the best possible clustering using 6 categories.

To find the best solution, the *k*-means algorithm was run with 25 random starting configurations. This was run 3 times to check it converged on the same clustering. This clustering is shown in Figure 5.5. This clustering corresponds fairly well with intuitive similarity perception and with the Euclidean distance that generated the space: all categories are contiguous except for the turquoise category. This category may have been influenced by noise from incomplete data: the bottom-right corner image included in this category was never compared to the images on its left or above, so the mean dissimilarity score may have been artificially high for these pairings. However, despite this slight noise, this category system serves as a reasonable baseline of the natural groups these images fall into on the basis of similarity.

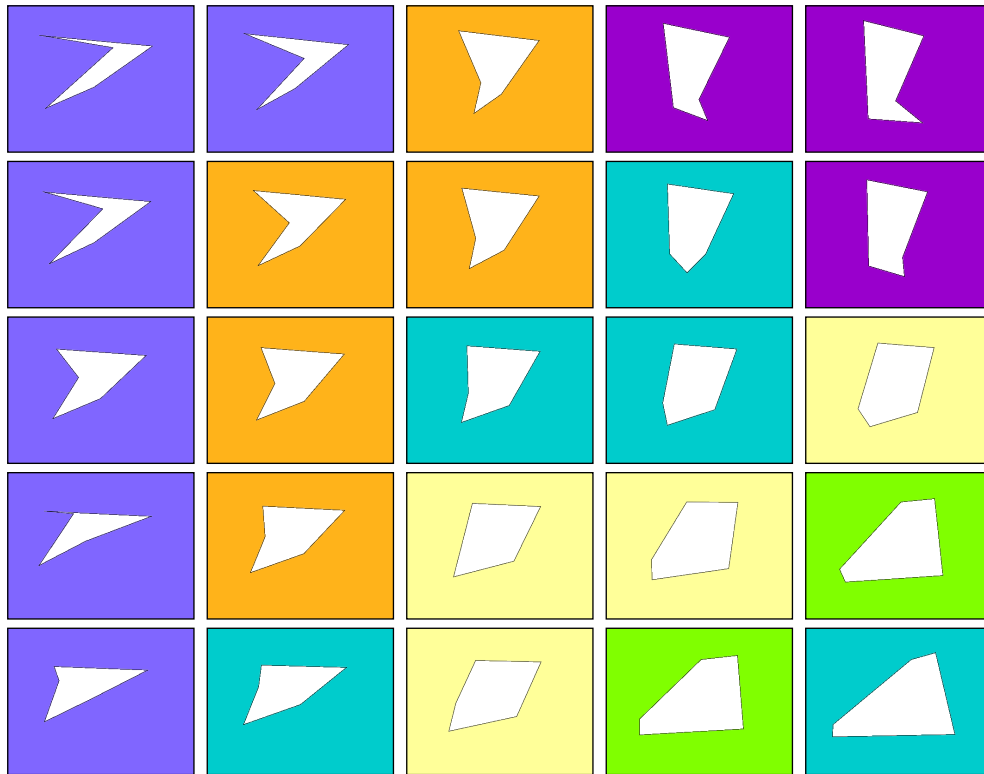


Figure 5.5: Category system most closely reflecting participants' pairwise similarity judgements, as assessed by k -means clustering. The turquoise category shows some evidence of perturbation by noise from missing data.

5.2.3.3 Characteristics of the similarity-based category system

Applying our dependent variables to the category system derived from participants' similarity ratings, we find the following:

Number of categories The category system divides the images into 6 categories, suggesting that the images naturally fall into 6 similarity-based groups.

Category convexity The system derived from participants' similarity ratings has a convexity score of 0.51. Perhaps surprisingly, the system is only about half as convex as it could be, given the number of categories and the distribution of images between categories (even though the *k*-means algorithm specifically aims to cluster category members as close as possible to category centres). Using this as a baseline, we can investigate whether communication tends to encourage category systems that are more or less convex.

Alignment Alignment with this similarity baseline system will be measured for all category systems produced in the experiment, in order to ascertain whether individual learning or communication encourages divergence from, or convergence to, category systems that reflect perceptual similarity.

5.3 Experiment 3

Experiment 3 compared category systems produced by individuals with category systems produced by communicating pairs.

5.3.1 Method

Participants were assigned to three conditions: two non-communicative conditions (whole-set non-communicative, sequential non-communicative) and a communicative condition. In the non-communicative conditions, participants divided the set of images from Experiment 2 into labelled categories on the basis of similarity. The whole-set and sequential non-communicative conditions differed in whether participants had access to the whole set of images when making their categorisation decisions (whole-set condition) or categorised images one at a time (sequential condition). In the communicative condition, pairs of participants played a communication game using the same

set of images, and produced labelled categories via the indirect process of selecting words to communicate each image to their partner.

5.3.1.1 Participants

Participants were 63 students at the University of Edinburgh (47 female, median age 24). 20 participants took part in the whole-set non-communicative condition. This condition took 15 minutes; participants were paid £2. 21 participants took part in the sequential non-communicative condition. One participant was subsequently excluded due to experimenter error. This condition took 30 minutes; participants were paid £3.50. 22 participants (randomly assigned into 11 pairs) took part in the communicative condition. This condition took an hour; participants were paid £7, and each member of the pair with the highest communication score was awarded a £10 Amazon voucher. One pair failed to complete the experiment within an hour and so were excluded from the analysis.

5.3.1.2 Stimuli

Stimuli were the set of 25 images drawn from a continuous space shown in Figure 5.1.

5.3.1.3 Labels

To control for any effects on participants' categorisations arising purely from the use of labels (Lupyan et al., 2007), words to label the categories were provided in both the non-communicative and communicative conditions. Lists of 25 CVCV nonsense words were generated by combining consonants and vowels randomly selected from the whole alphabet (e.g., zipi, gisa, wada). Since we expected that participants might use crossmodal associations between attributes of words and attributes of images in assigning labels (e.g. voiceless stops and spikiness, Nielsen & Rendall, 2011) we assigned the same wordlist to a yoked triple of two non-communicative participants and one communicative pair, so that in the analyses, any effects of a particular wordlist would apply equally across the conditions.

5.3.1.4 Procedure

Whole-set non-communicative condition Participants were presented with a randomised onscreen array of all 25 images and a set of words to label categories. They were instructed to label similar images with the same word and different images with

different words. Participants labelled categories by first clicking a word, then clicking an image or images to label them with this word. Participants could change their minds at any stage and either remove a label from an image or apply a different one. To avoid cueing the participants to produce a particular number of categories, only one word was initially shown on screen: participants could reveal new words at any time, up to a maximum of 25 words, by clicking a “new word” button, and were told that a) they could use as few or as many words as they wanted, and b) they did not have to use all the words they had revealed.

Communicative condition Participants completed the experiment in pairs, seated in separate cubicles and communicating via computer terminals. In a communication trial, one participant was designated as the sender and one as the receiver. The sender was presented with a randomised onscreen array of all 25 images, one of which was selected with a red box to indicate it was the target. The array of images was randomised anew for every trial in the experiment. The sender was also presented with one initial word. The sender could reveal a new word at any stage, by clicking a “new word” button, up to a maximum of 25 words. Any words they had revealed on a previous trial remained visible on their screen for all subsequent trials. The participant was instructed to choose a word that would help the receiver pick out the target from the array of images.

Once the sender had picked a word, the receiver was presented with a randomised onscreen array of all 25 images and the word the sender had chosen. The array of images was randomised anew for every trial in the experiment. The receiver was instructed to select the image the sender had wanted to communicate.

Once the receiver selected an image, both participants were presented with a feedback screen. The feedback screen showed the word the sender had used, the target image, the image the receiver had selected, the score for the trial, and the running total score for the whole experiment. The feedback screen was displayed for 4 seconds. The score for each trial was on a standardised scale calculated from the inverse Euclidean distance between the target and the image the receiver selected, from a minimum of 1 (for picking an image at the opposite corner of the space from the target) up to a maximum of 15 (for correctly picking the target).

After each communication trial the sender and the receiver swapped roles. The experiment consisted of 100 communication trials divided into 4 rounds. Each round featured the 25 images as targets in a randomised order. The randomised lists were bal-

anced such that each participant was the sender for every target image once in the first half of the experiment, and once in the second half. Halfway through the experiment, participants had the chance to take an optional break of up to 2 minutes.

Sequential non-communicative condition The sequential non-communicative condition was intended to mimic the sequential setup and amount of exposure to the stimuli in the communicative condition, but involved a single participant categorising stimuli purely on the basis of similarity (as in the whole-set non-communicative condition). In each trial, the participant was presented with a randomised onscreen array of all 25 images, one of which was selected with a red box to indicate it was the target. The participant was also presented with one initial word. The participant could reveal a new word at any stage, up to 25 words; any words they had revealed on a previous trial remained visible on their screen for all subsequent trials. Once the participant had picked a word for the stimulus, they were presented with the next trial. There were 100 trials in total, divided into 4 rounds, as in the communicative condition. Each round featured the 25 images as targets in a randomised order. As in the whole-set non-communicative condition, the participant was instructed to label similar images with the same word and different images with different words.

5.3.1.5 Dependent variables

Category systems were extracted from the experimental results as follows: In the whole-set non-communicative condition, the participant's category system consisted of the groups of images they labelled with the same word. In the sequential non-communicative condition, the participant's category system consisted of the groups of images they labelled with the same word in the fourth (i.e. final) round of the experiment. In the communicative condition, the pair's category systems were taken from the last two rounds of the experiment: since each participant labels half of all images in a given round, in order to get a complete category system per participant we have to look over the last two rounds, rather than the final round alone.

We quantified each category system according to a number of metrics.

Category measures The number of categories in each system, the convexity of the category system, and the alignment of pairs' category systems were measured as outlined in section 5.2.2 above. For the communicative condition, alignment was measured for pairs who communicated together. For the other two conditions, there were

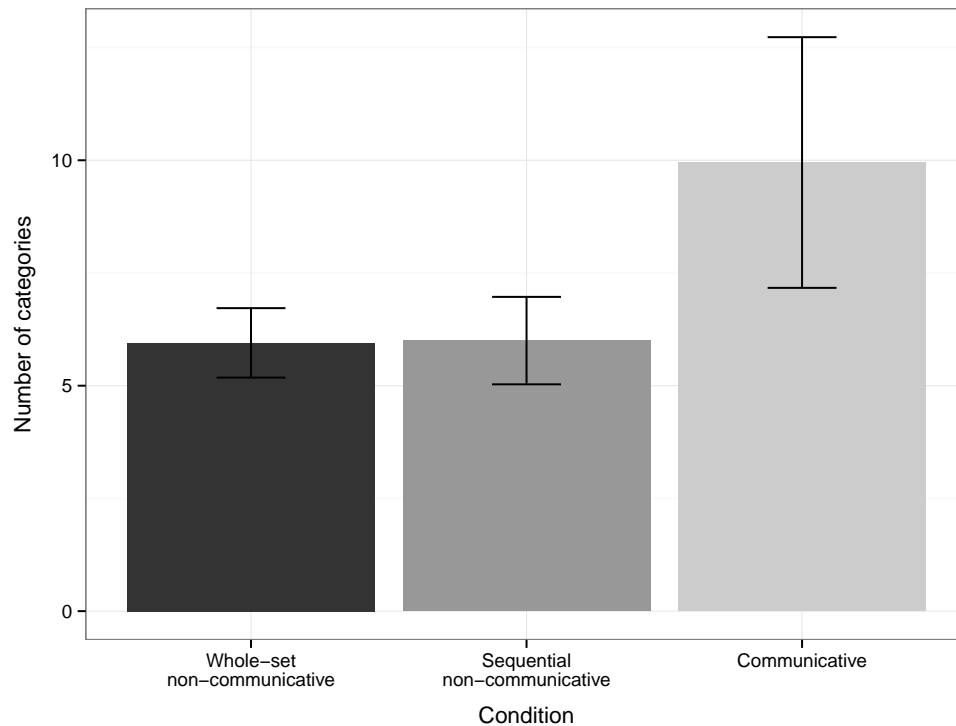


Figure 5.6: Number of categories used in the three conditions of Experiment 3. Error bars are 95% confidence intervals.

no ‘pairs’ as such; however, individuals who used the same wordlist were treated as pairs, and their alignment measured for the purpose of comparison.

Communicative success For the communicative condition, a further variable of interest was the extent to which pairs were successful in the communication game. Success was measured by the scores from 1-15 on each trial, summed per round and over the whole experiment.

5.3.2 Experiment 3 Results

5.3.2.1 Number of categories

Figure 5.6 shows the number of categories used by participants in each condition. Participants in the communicative condition used on average 9.95 labelled categories, 95% CI [7.17,12.73].⁶ This was more than in both non-communicative conditions, where

⁶All confidence intervals for the communicative condition were calculated based on standard errors and degrees of freedom for the average number for each pair, rather than counting each member of the pair separately (since these data points were not independent).

participants used a similar number of categories: whole-set non-communicators used a mean of 5.95 [5.18, 6.72], and sequential non-communicators a mean of 6.00 [5.03, 6.97]. The two non-communicative conditions only differed by 0.05 [-1.28, 1.38]. The differences between communicators and the two non-communicative conditions were larger: communicators used on average 4 more labelled categories than participants in the whole-set non-communicative condition [0.72, 7.28], and 3.95 categories more than participants in the sequential non-communicative condition [0.64, 7.26]. As the larger confidence interval for communicative participants suggests, communicating pairs varied substantially in how many labelled categories they used: *SD* for communicative participants = 3.88 [2.67, 7.08], for whole-set non-communicative participants = 1.07 [0.74, 1.95], for sequential non-communicators = 1.37 [0.94, 2.5].⁷ Levene's test for homogeneity of variance confirmed that this difference was significant, $F(2, 27) = 4.97$, $p = .01$. However, this variation came from differences between, rather than within, pairs; a paired-samples *t*-test found no significant difference in number of categories within each pair, $t(9) = 0.89$, $p = .40$.

Since the variance in the conditions was unequal and thus the data did not meet the assumptions for an ANOVA, a non-parametric test was run instead. A Kruskal-Wallis test confirmed a significant main effect of condition on number of categories, $H(2) = 10.87$, $p = .004$. Post-hoc Mann-Whitney tests were run on the pairwise differences between conditions, with the Bonferroni correction for multiple comparisons setting a significance threshold of $p = .017$. These tests found that the differences between the communicative condition and the other two conditions were both significant: for the whole-set non-communicative condition, $U = 11$, $p = .003$, and for the sequential non-communicative condition, $U = 14.5$, $p = .008$. However, the difference between the two non-communicative conditions was not significant, $U = 52$, $p = .91$. d_{umb} for the difference between communicators and whole-set non-communicators was 1.35, and for communicators versus sequential non-communicators was 1.3, suggesting a very large effect.

Both non-communicative conditions used a comparable number of categories to the similarity baseline shown in Figure 5.5. One-sample *t*-tests on each condition confirmed that while neither non-communicative condition was significantly different from the baseline category number of 6, $t(9) = -0.15$, $p = .89$, $t(9) = 0$, $p = 1$, the average number of categories in the communicative condition was significantly higher than the baseline, $t(9) = 3.22$, $p = .01$.

⁷These and all following 95% CIs on *SDs* calculated by the formula given in Sheskin (2011).

5.3.2.2 Communicative success

Average score in the first round of the experiment was 257 points [245, 269], and in the final round 295 points [272, 318]. Communicators improved by an average of 38 points from the first to the last round [18, 58]. Chance performance for each round was 245, suggesting that pairs started by predominantly guessing, but developed and improved strategies for communication over the course of the experiment. A linear trend ANOVA found that communicative success increased significantly over rounds in the experiment, $F(1,9) = 4.13$, $p = .05$. d_{unb} was 1.38, suggesting a very large improvement in communicative success over rounds.

For validity of comparison with category measures, the following analyses are of the final two rounds of communication only. Participants' success in the last two rounds of communication was significantly above chance, $t(9) = 4.56$, $p = .001$. However, pairs' strategies varied in their success (as shown by the large confidence interval on final-round scores). Interestingly, it was not the case that pairs who used more categories were more successful: the correlation between average number of categories and communicative success in the last two rounds was negative and non-significant, $r = -0.30$ [-0.78, 0.41]⁸, $p = .41$. See section 5.5.1 below for further discussion of factors affecting communicative success.

5.3.2.3 Category convexity

The category systems of whole-set non-communicators had an average convexity score of 0.77 [0.72, 0.82]. Sequential non-communicators had an average convexity of 0.65 [0.58, 0.72], and communicators, 0.57 [0.41, 0.73]. Communicators' convexity indices were lower by 0.2 [0.04, 0.36] than whole-set non-communicators, and by 0.08 [-0.08, 0.24] than sequential non-communicators. The sequential non-communicators had convexity indices 0.12 [0.04, 0.20] lower than their whole-set counterparts. Communicators had the least convex category systems of the three conditions (Figure 5.7). Again, communicators varied more widely than participants in the other two conditions: SD for whole-set non-communicators = 0.07 [0.05, 0.13], for sequential non-communicators = 0.09 [0.06, 0.16], and for communicators = 0.22 [0.15, 0.40]. Levene's test found that this difference in variance was significant, $F(2,27) = 4.81$, $p = .02$. Since the variance in the conditions was unequal and thus the data did not meet

⁸This and all following confidence intervals on r are calculated using the formula given in Cumming (2012).

the assumptions for an ANOVA, a non-parametric test was run instead. A Kruskal-Wallis test confirmed a significant main effect of condition on category convexity, $H(2) = 8.98$, $p = .01$. Post-hoc Mann-Whitney tests were run on the pairwise differences between conditions, with the Bonferroni correction for multiple comparisons setting a significance threshold of $p = .017$. These tests found that the difference between the two non-communicative conditions was significant, $U = 93$, $p < .001$. d_{unb} for this difference was 1.43, suggesting a very large effect. However, due to its high variance the communicative condition was not significantly different from either the whole-set non-communicative condition, $U = 75$, $p = .06$, or the sequential non-communicative condition, $U = 56$, $p = .68$.

Notably, all of these convexity values were numerically above the value of 0.51 for the similarity baseline category system. One-sample Mann-Whitney tests on each condition (correcting for multiple comparisons to set a significance level of $p = .017$) found that both non-communicative conditions had mean convexity significantly above 0.51: $U = 55$, $p = .002$ (whole-set non-communicators), $U = 53$, $p = .006$ (sequential non-communicators). However, at 0.57, communicators' mean convexity was not significantly higher than the baseline system, $U = 36$, $p = .43$.

5.3.2.4 Category system alignment

Surprisingly for the hypothesis, communicative pairs' category systems were the least aligned out of the three conditions. Whole-set non-communicators were the most aligned, with a mean adjusted Rand index of 0.48 [0.37, 0.59], following by sequential non-communicators with a mean of 0.33 [0.22, 0.44]. Communicators were the least aligned, with a mean of 0.24 [0.08, 0.40]. The difference between communicators and whole-set non-communicators was 0.24 [0.05, 0.43], or a quarter of the possible range of the Adjusted Rand index (0 to 1). The difference between communicators and sequential non-communicators was smaller, at 0.09 [-0.09, 0.27]. The whole-set non-communicators were on average 0.15 more aligned than the sequential non-communicators [0.00, 0.30]. A one-way ANOVA found a significant main effect of condition, $F(2, 27) = 4.28$, $p = .02$. Tukey post-hoc tests found that this effect was driven by the difference between communicators and whole-set non-communicators, $p = .02$. d_{unb} for this difference was 1.14, suggesting a very large effect. Neither of the other two pairwise comparisons was significant: $p = .19$ for whole-set vs. sequential non-communicators, $p = .52$ for communicators vs. sequential non-communicators.

To test the hypothesis that participants within a pair were more aligned than par-

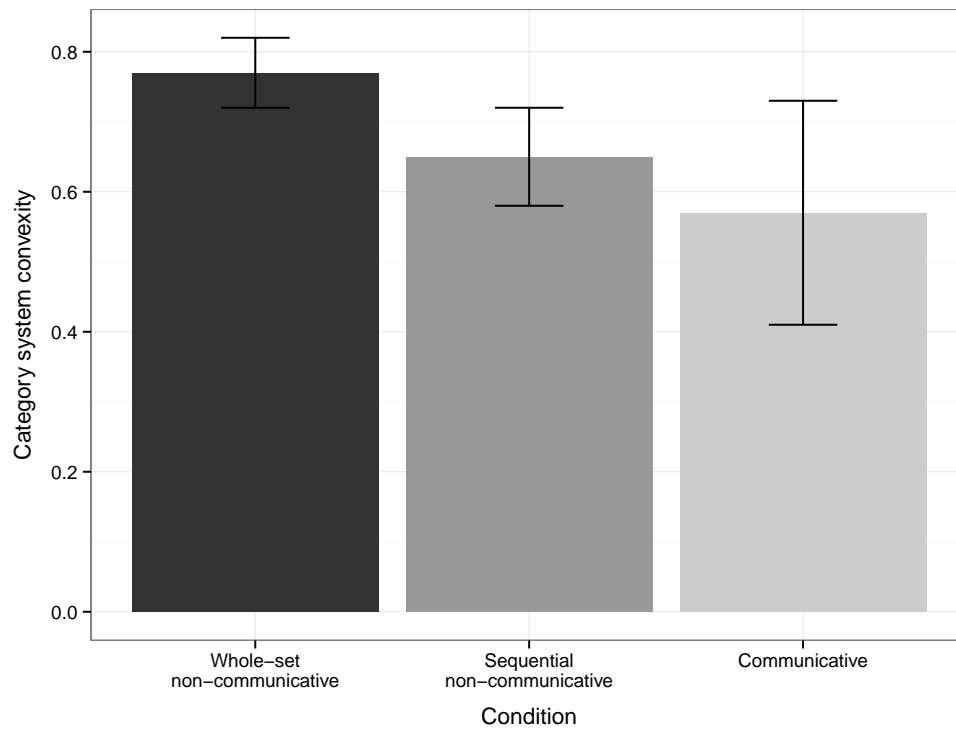


Figure 5.7: Difference in category system convexity over the three conditions of Experiment 3, measured by average proportion of neighbors in the same category. Error bars are 95% confidence intervals.

ticipants who were not paired with each other, a Monte Carlo simulation was run to find the average alignment across pairs. On each round, participants' systems within a condition were randomly paired, and their alignment scores averaged. Monte Carlo simulations were run for 10000 rounds. This produced a distribution of mean across-pair alignment scores. The proportion of times the veridical mean was higher than or equal to the simulated mean was then taken as a measure of the probability of across-pair alignment being as high as or higher than the veridical alignment within pairs.

Participants in both non-communicative conditions were not significantly more aligned within pairs than general across-pair levels of alignment. For whole-set non-communicators, the probability of across-pair alignment being as high as veridical within-pair alignment was .40. For whole-set non-communicators, the probability of across-pair alignment being as high as within-pair alignment was .78. In contrast, for communicators, the probability of across-pair alignment being as high as within-pair alignment was .02, suggesting that within-pair alignment was reliably higher. Figure 5.8 shows the mean across-pair alignment values and the veridical alignment values for each condition.

The category systems resulting from the three conditions were also compared to the similarity baseline to determine which condition had greater alignment with this baseline. Whole-set non-communicators had an average alignment of 0.36 [0.32, 0.40] with the baseline system, while sequential non-communicators' alignment with the baseline system was 0.34 [0.30, 0.38]. The difference between these was negligible, 0.02 [-0.04, 0.08]. Communicators had an average alignment of 0.21 [0.15, 0.27] with the baseline system. This was 0.15 [0.08, 0.22] lower than whole-set non-communicators, and 0.13 [0.06, 0.20] lower than sequential non-communicators. A between-subjects ANOVA confirmed that there was a main effect of condition, $F(2, 57) = 10.94$, $p < .001$. Post-hoc Tukey tests found that the difference between the communicative condition and both non-communicative conditions was significant, $p < .001$ for the comparison with the whole-set condition ($d_{umb} = 1.36$) and $p = .001$ for the comparison with the sequential condition ($d_{umb} = 1.10$); however, there was no significant difference between the two non-communicative conditions, $p = .84$. As Figure 5.9 shows, alignment with the baseline system was generally low, suggesting that the task of sorting into labelled categories imposes fundamentally different pressures from judging pairwise similarity.

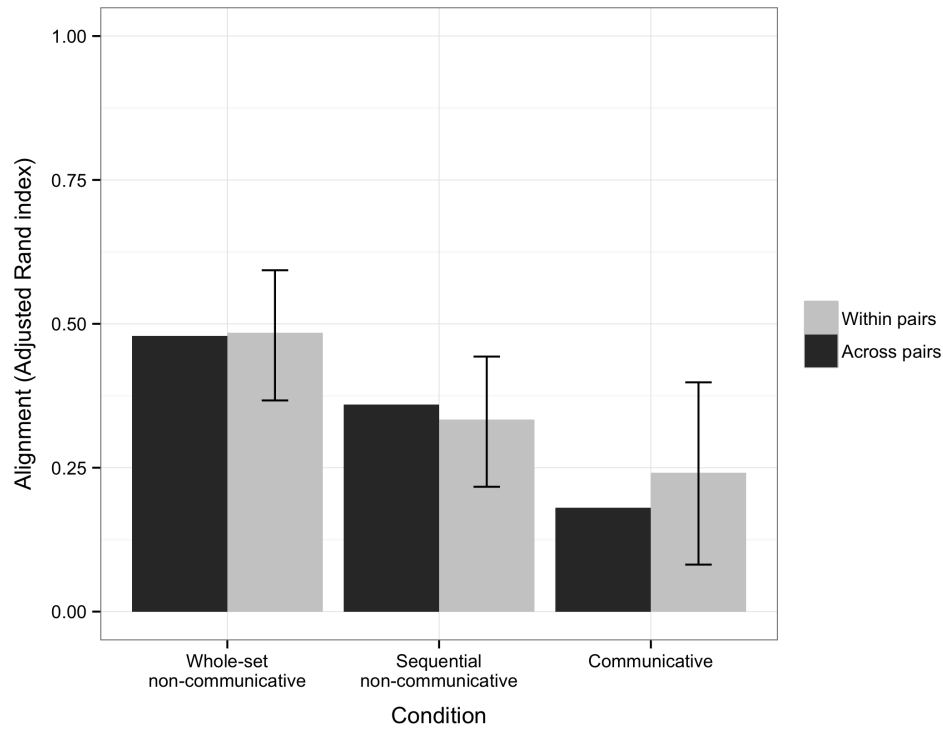


Figure 5.8: Average alignment of category systems in each condition of Experiment 3, measured by the Adjusted Rand index. Light bars show veridical alignment values within pairs, while dark bars show the mean alignment across pairs, generated by Monte Carlo simulations. Error bars are 95% confidence intervals. Across-pair mean alignment values are generated from simulation and hence do not have confidence intervals; see text for statistical comparison of mean within-pair alignment values to the simulated across-pair distribution.

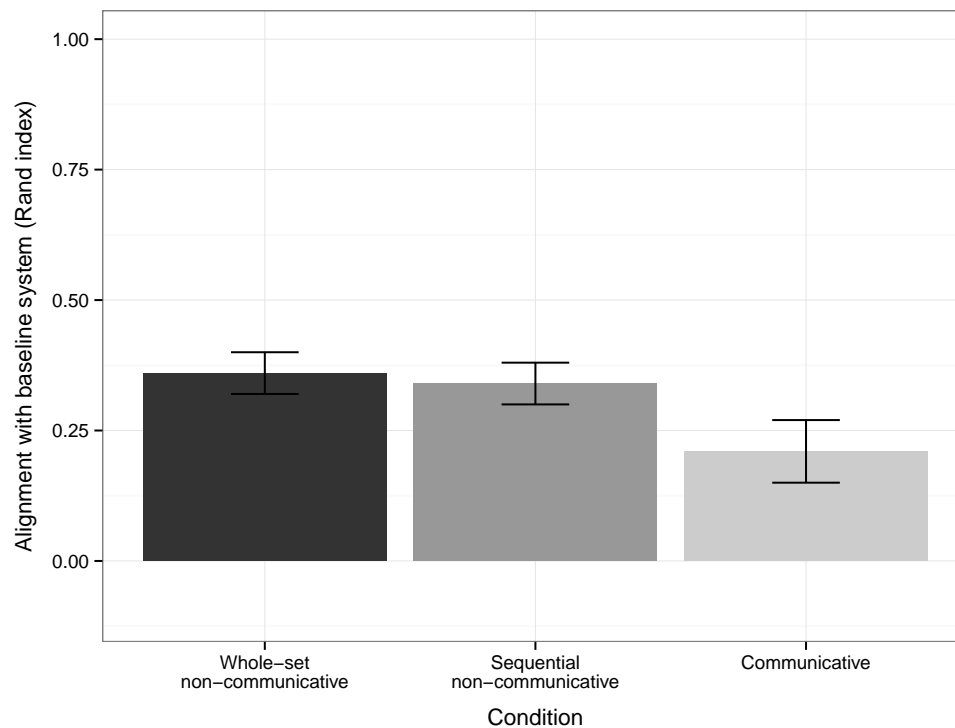


Figure 5.9: Average alignment of category systems in each condition with the similarity baseline category system shown in Figure 5.5. Error bars are 95% confidence intervals.

5.3.3 Experiment 3 Discussion

Communicators tended to use more categories than participants in either of the non-communicative conditions, supporting the hypothesis that communication encourages greater expressivity. However, communicative pairs varied widely in the number of categories they used, and using more categories was not correlated with higher success. In practice, the precision advantage of more categories seems to be offset by disadvantages for memory and coordination: increasing the number of categories is only useful if a) the participant can cope with the increased memory burden and b) the participant's partner can also acquire and remember the additional categories. This links back to the point made in section 3.5.3 of Chapter 3: expressivity is limited by what an interlocutor is capable of inferring, given the communicative context and the patterns of word reuse in the rest of the lexicon. Here, the communicative context is impoverished and does not provide additional clues beyond the lexicon, and the lexicon may not be completely shared, thus limiting the expressive potential even of a system with a large number of potential distinctions.

Interestingly, even though the non-communicative participants had no reason to

converge on a particular number of categories, they were more consistent than the communicative participants, even though the conditions involved two very different categorisation methods (sequential and whole-set). Given the match between this number and the baseline similarity system, this reinforces the idea that perceptually, the set of images fell into approximately 6 categories.

Whole-set non-communicators showed high levels of convexity, i.e., their categories consisted of images which clustered together in the original Euclidean space (even though the images were presented in a shuffled arrangement). This supports the idea that, at least in this perceptual space, human categorisation biases favour convex categories. However, sequential non-communicators showed lower levels of convexity, suggesting that the memory constraints imposed by having to categorise images sequentially either disrupted participants' perception, or simply made them forget which category they had placed particular images in. Communicators' categories were on average less convex still, although the variability between pairs was substantial. However, it is surprising that communication did not lead to convex categories, given the advantages of these categories for the communicative task in the experiment, as well as the hypothesised advantages of these structures for real language (Gärdenfors, 2000). The relation of these values to the similarity baseline system is also interesting. Both non-communicative conditions produced category systems with significantly higher convexity than the baseline system, suggesting that the act of placing stimuli into categories encourages more convex groupings than arise from similarity alone. Related to this, Warglien & Gärdenfors (2011) suggest that we should expect to find convex conceptual representations even in cases where the world's structure is not necessarily convex. However, the authors argue for this as a specifically communicative bias: 'seeing a non-convex world with convex spectacles might be a peculiar bias arising from selective pressures towards effective communication' (Warglien & Gärdenfors, 2011, p. 2178). The results reported here suggest that this bias may be more general to individual categorisation. At least in the communicative condition in this experiment, selective pressures towards convex categories arising from communication do not appear to have a strong effect: communicators are the only participants to produce category systems which are not significantly more convex than those that arise from similarity alone.

For alignment, the picture is also surprising. Communicators were the least aligned of the three conditions, despite the 'pairs' of individuals from the other two conditions having no direct interaction, and therefore no mechanism for convergence other than

shared categorisation biases. Taking whole-set non-communicators as a baseline, individual participants in this task have an alignment of about 50% over chance in their categorisations of the stimulus space; however, being made to categorise images sequentially instead of all at once (i.e. not ‘seeing’ the categories they are constructing, but having to commit them to memory) disrupts this alignment in the sequential non-communicative condition. This has a precedent in previous modelling work: for example, Steels & Belpaeme (2005) found that when their agents built up categories individually via a series of discrimination games, their category systems were poorly aligned as a result of their different experiential histories (Steels & Belpaeme, 2005, p. 478); A. Smith (2003a) also found poor alignment for agents who had different discrimination histories, in the absence of a structured world that channeled meaning distinctions to appropriate dimensions. Likewise, the sequential non-communicators in Experiment 3 encounter and categorise the images in different orders, which may lead to idiosyncratic effects on their categorisations. Communication may work to disrupt this alignment still further, perhaps because of the additional noise caused by participants having the extra burden of learning their partner’s category system at the same time as constructing their own. Another possibility is that participants continually accommodate to each other in a back-and-forth process, resulting in a lack of stability for each individual’s category system (this back-and-forth accommodation process was mentioned by several participants in post-experiment questionnaires). However, the smaller size of the difference between communicators and sequential non-communicators means it is hard to be certain about the negative effects of communication over and above the memory constraints of sequential labelling.

To summarise: as predicted in the Introduction, communicators’ systems are generally more expressive than those produced by individuals (although levels of expressivity vary widely between pairs). However, the results seem to go against the other hypotheses. Communicators’ category systems did not have high levels of convexity, despite theoretical arguments that convex categories are more optimal for communication. Communicators are also the least aligned of the three conditions, despite the interactive alignment account predicting that communication should lead to alignment of individuals’ representations.

One possible explanation for the low level of convexity and alignment in the communicative condition is the unusual pressure of building a communicative system from scratch. Communicators are forced to learn about the structure of the space at the same time as coordinating with their partner, perhaps creating a situation where coordinating

conventions for specific images is a more accessible solution than coordinating on a fully structured and aligned category system (see section 5.5.1 for further discussion). However, in the real world, humans learn the culturally transmitted category system of their native language. Previous work has shown that cultural transmission of behaviours leads those behaviours to become more structured (Kirby et al., 2008; Reali & Griffiths, 2009; Smith & Wonnacott, 2010; Theisen-White et al., 2011), i.e. exhibiting patterns and regularities which facilitate learning. It could be that these learning-enabled structures need to be in place before communication can work to optimise and align them further.

In Experiment 4, we tested this hypothesis by simulating cultural transmission of category systems using iterated learning. The general methodology for iterated learning is as in Experiment 1, described in Chapter 4. Here, a ‘generation’ consists of a pair, who 1) learn a category system, 2) play the same communication game as in Experiment 3, and 3) pass their final category system on as learning input to the next pair in a chain. The aim was to discover how learning and communication interact to shape category structures. In particular, we were interested in finding out whether cultural transmission would lead to increasing structure in category systems, and whether this would alter the effect of communication, providing common ground from which communication could work to make category structures more convex, more aligned, and less variable.

5.4 Experiment 4

5.4.1 Method

Pairs of participants learned a labelled category system, then played the same communication game as in Experiment 3 (with small methodological differences: see below for details and justification). The first-generation pair in each chain learned a system where every image was in its own category. Subsequent generations learned the category system produced by a randomly selected member of the previous pair at the end of their communication phase. The category systems the participants produced at the end of learning and at the end of communication in each generation were analyzed.

5.4.1.1 Participants

Participants were 90 students at the University of Edinburgh (62 female, median age 24). 2 participants were excluded due to experimenter error, 4 due to networking issues, and 4 due to not completing the experiment within the allotted time. This left 80 participants, randomly assigned into 8 chains of 5 generations, with each generation consisting of a pair. The experiment took 90 minutes. Participants were paid £10, and each member of the pair with the highest communication score was awarded a £10 Amazon voucher.

Stimuli were identical to Experiments 2 and 3. Labels were constructed within the same parameters as the labels in Experiment 3, with each pair in Experiment 4 using a different wordlist. This was to avoid associations between particular words and images having a systematic effect within chains or generations.

5.4.1.2 Procedure

The experiment consisted of two phases: a learning phase and a communication phase. Participants completed the experiment in pairs, seated in separate cubicles and communicating via computer terminals.

Learning phase During the learning phase, the participant's task was to learn a language consisting of words that applied to the 25 images in the set. Participants were informed that their partner was learning the same language. For generation 1 participants, this language consisted of 25 words, with one applying to each image: i.e., each image belonged to its own category. For subsequent generations, the target language was the category system produced by the previous generation during their communication phase; however, the wordlist was substituted for a new one. That is, the groups of images given the same label were preserved from the previous generation, but a new word was applied randomly to each category. This inter-generational shuffling process was intended to minimise the cumulative impact of sound-symbolic biases favoring certain image-word pairings.

The method for displaying words on screen was changed from Experiment 3. In Experiment 3, one word was displayed initially and the participant could click to reveal more. However, this method would be frustrating for participants undertaking a learning task, since they might have to click repeatedly to reveal the correct word. Instead, 30 words were displayed and remained onscreen for all trials. 30 words were chosen

to make it clear to participants that there were more words than images and thus they did not have to use all the words. Participants within a pair had the same wordlist. The order of words as presented on screen was randomised independently for each participant, but remained constant for a given participant throughout the learning and communication phases of the experiment.

On each learning trial, the participant was presented with the 30 words and a randomised onscreen array of all 25 images, one of which was selected with a red box to indicate it was the target. The array of images was randomised anew on each experimental trial. The participant was instructed to click the word for the selected image. Once the participant had clicked a word, they were presented with a feedback screen. The feedback screen told them if their word choice was correct or incorrect, and displayed the target image and the correct word for 4 seconds before moving to the next trial. Thus, learning was implicit: initially, participants were forced to guess, but they had the opportunity to gradually acquire the category system via feedback.

The learning phase ran for 100 trials, divided into 4 rounds. Each round featured the 25 images as targets, in a randomised order. At the end of each round, the participant who had finished first was shown a holding screen until their partner had also finished that round, at which point both proceeded to the next round. Halfway through the learning phase, participants had the chance to take an optional break of up to 2 minutes.

Participants' category systems at the end of learning were defined as the groups of images they labelled with the same word in the final round of the learning phase.

Communication phase Participants were instructed that they were going to play a communication game with their partner, using words to communicate the identity of the selected images. They were not specifically instructed to use the language they had learned. The procedure for the communication phase was identical to the communicative condition of Experiment 3, except for the presentation of words onscreen to the sender, which mirrored the learning phase of Experiment 4: i.e. all 30 words were provided from trial 1, rather than the participant clicking to reveal new words as required.

Iteration After each run of the experiment, one of the pair was randomly selected to provide the input for the next generation. That participant's category system produced during the last two rounds of communication became the target system for learning by the next pair in a chain.

5.4.1.3 Dependent variables

Number of categories, category system contiguity, alignment, and communicative success were measured as in Experiment 3. The following additional measure was also applied to the iterated experiment:

Learnability An additional variable of interest in Experiment 4 was whether participants were able to acquire the input category system during the learning phase, and to what extent this system remained stable through the communication phase. Transmission error, i.e., the amount a participant's category system differed from the system they learned, was assessed by two measures: 1) the proportion of images assigned the wrong word in the final round of the learning and communication phases, and 2) dissimilarity between the category systems participants were trained on and those they produced in the learning and communication phases (measured by $1 - \text{the Adjusted Rand index}$).

5.4.2 Experiment 4 Results

5.4.2.1 Learnability

Figure 5.10 shows error over generations, measured by the proportion of images participants labelled incorrectly (words-based error).

Words-based error in the final round of learning decreased over generations in the experiment. Error was 0.81 [0.72, 0.90] in generation 1, dropping to 0.4 [0.29, 0.51] by generation 5. The decrease in error over generations was 0.41 [0.28, 0.54]. A linear trend ANOVA found that this decrease was significant, $F(1, 75) = 37.4, p < .001$. d_{unb} for the drop from generation 1 to 5 was 2.26, suggesting a very large effect.

Words-based error in the final round of communication also decreased across generations. Error by the end of communication was 0.78 [0.69, 0.87] in generation 1, dropping to 0.3 [0.16, 0.44] by generation 5. The decrease in error from generation 1 to 5 was 0.48 [0.33, 0.63]. A linear trend ANOVA found that this decrease in error was significant, $F(1, 35) = 24.7, p < .001$. d_{unb} for the drop from generation 1 to 5 was 3.21, suggesting a very large effect. These error scores were slightly but consistently lower than the error scores at the end of learning, suggesting that communicating with a partner who had been trained on the same system allowed participants to pool what they remembered, leading to more word-image correspondences being recalled by the

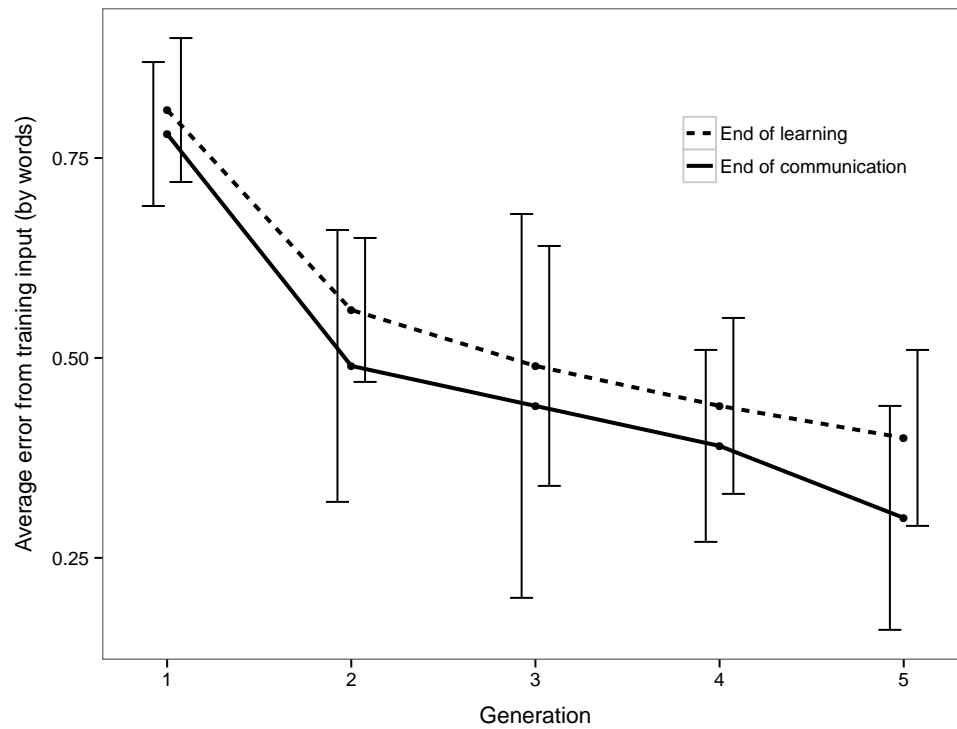


Figure 5.10: Graph showing decrease in error over generations in Experiment 4, measured by proportion of images participants gave the wrong label in the final round of the learning phase (dotted line) and the communication phase (solid line). Error bars are 95% confidence intervals.

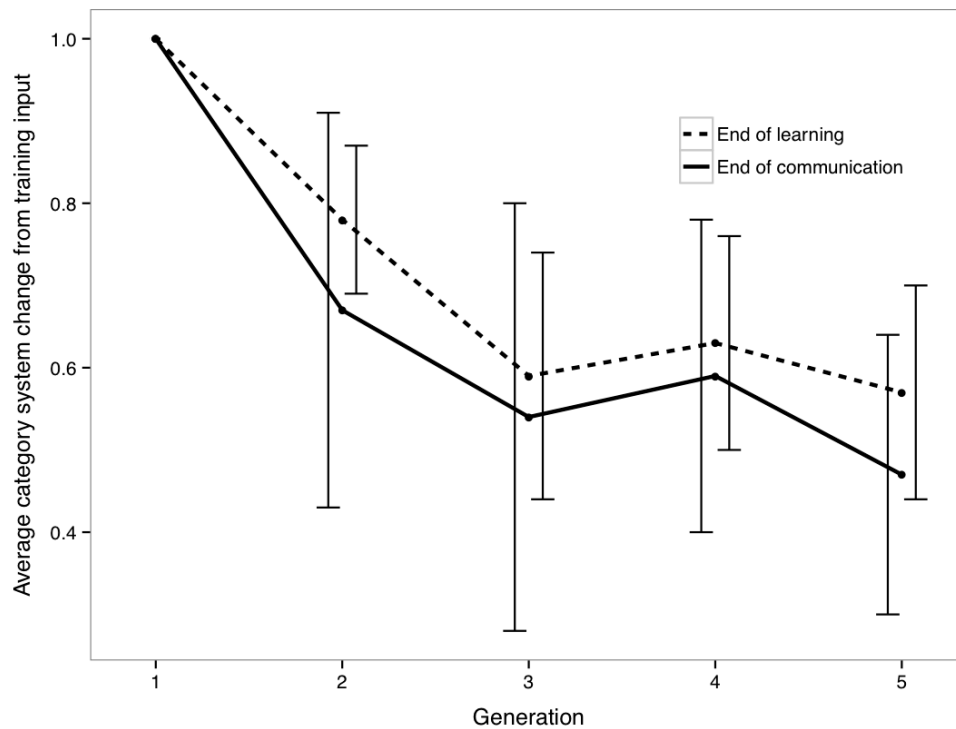


Figure 5.11: Graph showing decrease in the amount of category system change over generations in Experiment 4, measured by $1 - \text{alignment}$ (Adjusted Rand index) between the system participants were trained on and the system they produced in the final round of the learning phase (dotted line) and the communication phase (solid line). Error bars are 95% confidence intervals.

end of communication.

Figure 5.11 shows amount of change in category systems at each generation, based on the dissimilarity (i.e. $1 - \text{alignment}$) of participants' final category systems from the category system they were trained on.

Category system change at generation 1 was always 1 by definition. Since the training system had one image per category, and therefore every pair of images was in different categories, there was no difference between chance alignment and veridical alignment between this and any other system. Therefore the corrected-for-chance alignment measure was always 0, and dissimilarity was always 1.

Category system change in the final round of training was 0.78 [0.69, 0.87] in generation 2, dropping to 0.57 [0.44, 0.70] by generation 5. There was a decrease of 0.21 [0.06, 0.36] from generation 2 to generation 5. A linear trend ANOVA found that this decrease was significant, $F(1, 75) = 36.4$, $p < .001$. d_{umb} for this drop was 0.98, suggesting a large effect.

Category system change in the final round of communication was 0.67 [0.43, 0.91] in generation 2, dropping to 0.47 [0.30, 0.64] by generation 5. The amount of change decreased by 0.2 [-0.06, 0.46] from generation 2 to generation 5. Again, a linear trend ANOVA found this trend was significant, $F(1,35) = 19.7$, $p < .001$. d_{umb} for the change from generation 2 to 5 was 0.78, suggesting a large effect. As with words-based error, category system change after communication was consistently lower than error after learning, suggesting that communication helped pairs converge closer to the category structure of the system they were trained on.

5.4.2.2 Number of categories

Figure 5.12 shows the change in number of categories over generations of learning and communication. At the end of learning, pairs had an average of 15.56 [13.79, 17.33] categories in generation 1, dropping to 7.5 [6.20, 8.80] by generation 5. The number of categories at the end of learning decreased by 8.06 [6.04, 10.08] from generation 1 to 5.

The average number of categories in pairs' systems at the end of the communication phase was also measured. In generation 1, by the end of communication, pairs were using an average of 12.5 categories [9.83, 15.17]. This was on average 2.55 [-1.07, 6.17] more than in the communicative condition from Experiment 3, but as the large confidence interval shows, pairs varied widely: $SD = 3.2$ [2.12, 6.51]. This also represented a drop of 3.06 categories from the end of the learning phase [0.55, 5.57]. The number of categories at the end of communication dropped over generations; however, for generations 2, 3, and 4, the average number of categories increased between the end of learning and the end of communication. By the end of communication in generation 5, the number of categories had dropped to 7.5 [6.32, 8.68]. In this generation there was no increase in number of categories from learning to communication: 95% CI for this difference [-0.64, 0.64]). This final number was 1.55 [0.24, 2.86] more categories than whole-set non-communicators from Experiment 3. A between-groups t -test found that the number of categories in generation 5 of Experiment 4 was significantly higher than the number of categories produced by whole-set non-communicators from Experiment 3, $t(16) = 2.66$, $p = .02$. This was also significantly higher than the similarity baseline number of 6, $t(7) = 3$, $p < .05$. The overall decrease in number of categories at the end of communication from generation 1 to 5 was 5 [2.23, 7.78], $d_{umb} = 1.80$. The variability between pairs in number of categories also dropped over generations, SD at generation 5 = 1.41 [0.93, 2.87].

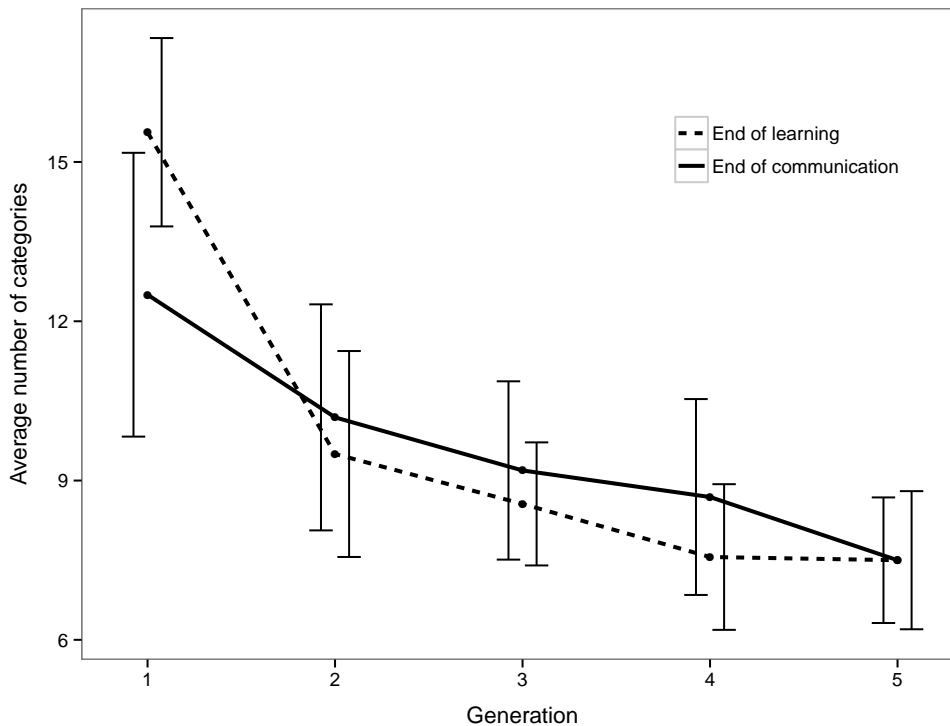


Figure 5.12: Change in number of categories over generations in Experiment 4, at the end of the learning phase (dotted line) and the communication phase (solid line). Error bars are 95% confidence intervals.

A two-way mixed ANOVA with Generation as a between-subjects factor and experimental phase (learning versus communication) as a within-subjects factor found a significant linear trend for Generation on the number of categories, $F(1, 35) = 47.7$, $p < .001$. d_{unb} for the drop over generations at the end of learning was 4.08, suggesting an extremely large effect. d_{unb} for the corresponding drop over generations at the end of communication was 2.02, suggesting a smaller (because of the lower number of categories post-communication in generation 1) but still substantial effect.

There was no main effect of experimental phase (learning versus communication), $F(1, 35) = 0.20$, $p = .66$. However, there was a significant interaction of Generation and experimental phase, $F(4, 35) = 7.18$, $p < .001$, reflecting the fact that the number of categories dropped between learning and communication in generation 1 ($d_{unb} = 1.00$), but increased in all other generations. d_{unb} values for the increases in number of categories between learning and communication in generations 2, 3 and 4 ranged from 0.25 in generation 2 to 0.51 in generation 4, suggesting a small to medium effect of communication in increasing the number of categories participants used.

5.4.2.3 Communicative success

Performance in generation 1 was broadly similar to Experiment 3, but with a higher initial score and smaller improvement over rounds. Generation 1 participants scored an average of 277 points [254, 300] in the first round of communication, an increase of 20 points [-3.94, 43.94] over the round 1 scores of communicators in Experiment 3. By round 4, generation 1 participants scored an average of 302 [276, 328], which was a difference of only 7 points [-24.74, 38.74] from the round 4 scores of communicators in Experiment 3. The increase from the first to the final round for generation 1 participants in Experiment 4 was 25 points [2.98, 47.01]. In subsequent generations, improvement over rounds in the communication game was generally smaller, ranging from 9 points [-5.76, 23.76] in generation 2 to 12 points [-0.13, 24.13] in generation 5, suggesting that for these generations strategies were more stable from the start to the end of communication.

For validity of comparison with category measures, all following analyses are of the final two rounds of communication only.

Communicative success improved over generations. Success scores in the last 2 rounds increased from 595 [552, 638] in generation 1 to 649 [627, 671] in generation 5; the average increase in score in the last 2 rounds from generation 1 to 5 was 54 points [11, 97]. Chance success, as before, was 245 per round and therefore 490 for 2 rounds; a one-sample t -test found that even in generation 1, communicators were performing significantly above chance, $t(7) = 5.72$, $p < .001$. This suggests that from the first generation, pairs were already developing functional category systems. However, the communicative efficiency of the systems still improved over generations, $F(1, 35) = 7.61$, $p = .009$. d_{umb} for the increase from generation 1 to 5 was 1.22, suggesting a very large effect.

This increase in communicative success is particularly notable given that, as detailed in the previous section, the average number of categories used by pairs was decreasing over generations. Systems with fewer categories might be expected to be less communicatively successful, since they are less expressive. To explore this, communicative success scores in the last two rounds were divided by each pair's average number of categories, to correct for the expected degree of precision. The result is a smooth trend upwards (see top panel of Figure 5.13); i.e., when the decreasing number of categories is taken into account, communicative success increased linearly over generations.

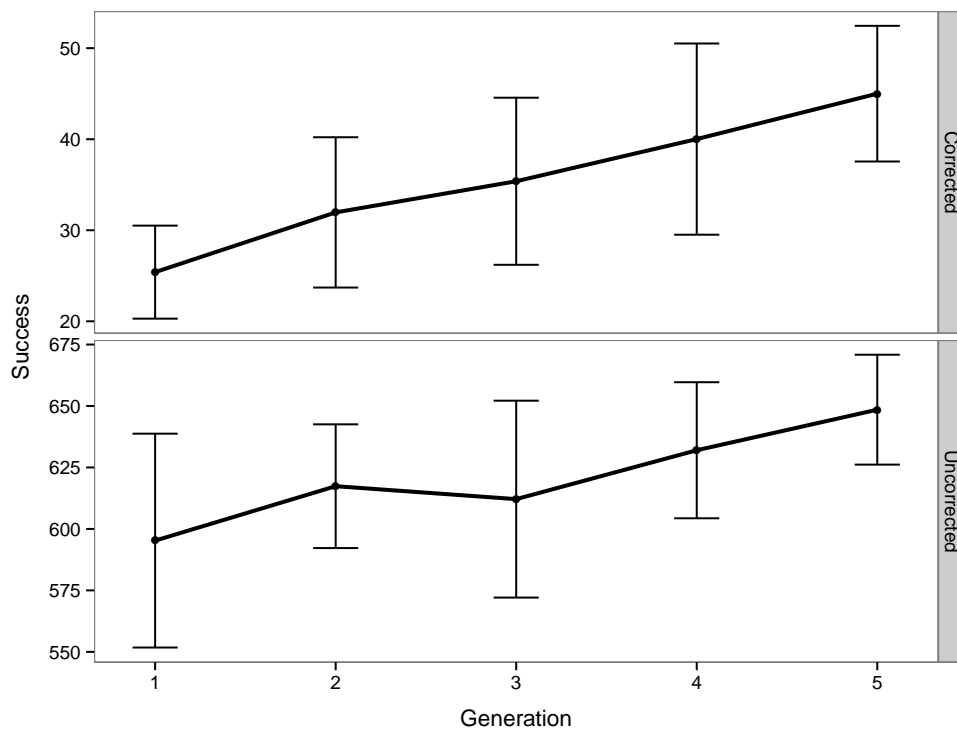


Figure 5.13: Graph showing success scores in the last 2 rounds of the communication game over generations in Experiment 4. The top panel shows success scores corrected for number of categories (i.e., communicative success over the last 2 rounds divided by the average number of categories used by the pair during those rounds). This number equates to average points gained per category in the system. The bottom panel shows raw success scores. Error bars are 95% confidence intervals.

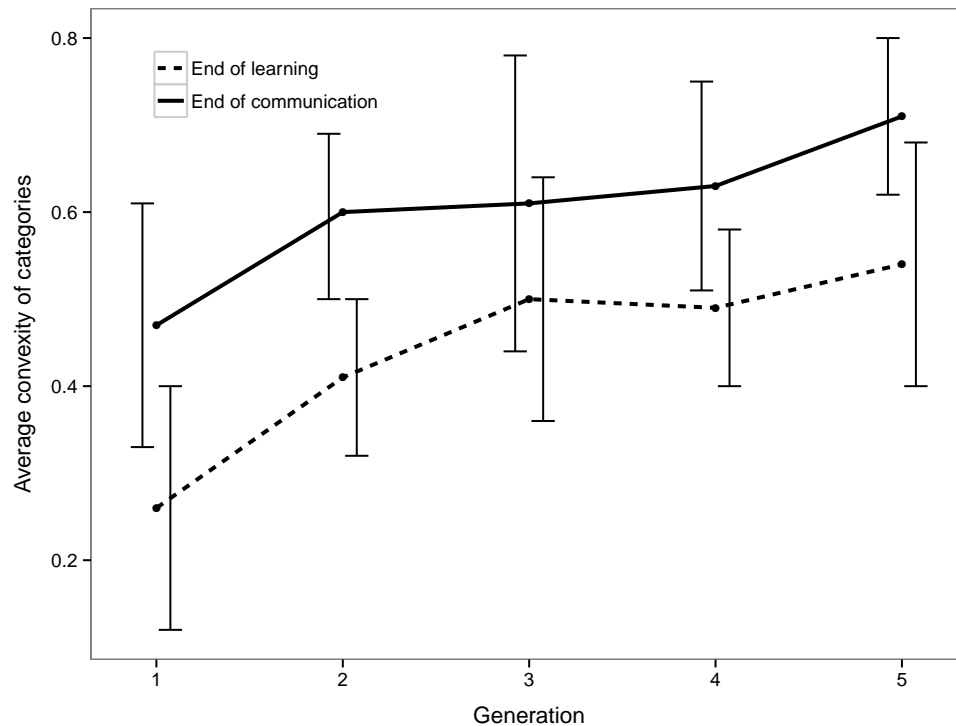


Figure 5.14: Graph showing change in pairs' category system convexity over generations in Experiment 4. Dotted line shows convexity post-learning, solid line convexity post-communication. Error bars show 95% confidence intervals, with standard errors adjusted to reflect only between-subjects differences.

5.4.2.4 Category convexity

Category system convexity increased over generations. Figure 5.14 shows the change in convexity over generations for category systems at the end of learning and at the end of communication. In generation 1, convexity at the end of learning was 0.26 [0.12, 0.40], and after communication 0.47 [0.33, 0.61]. A t -test found that this value of 0.47 was not significantly different from the communicators in Experiment 3, whose value was 0.57 on average, $t(16) = -1.12$, $p = .28$. By generation 5, convexity was 0.54 [0.40, 0.68] post-learning and 0.71 [0.62, 0.80] post-communication. The increase between generation 1 and 5 post-learning was 0.28 [0.09, 0.47], while the increase between generation 1 and 5 post-communication was 0.24 [0.08, 0.40]. By the end of communication in generation 5, the average convexity value of 0.71 was statistically comparable to the value of 0.77 from whole-set non-communicators in Experiment 3, $t(16) = -1.25$, $p = .23$.

The convexity of pairs' systems increased during the communication phase at each

generation. The increase between the end of learning and the end of communication in generation 1 was 0.2 [-0.01, 0.41], in generation 2 0.19 [0.02, 0.36], in generation 3 0.11, [-0.02, 0.26], in generation 4 0.14 [0.02, 0.26], and in generation 5 0.17 [0.05, 0.29].

A mixed ANOVA with Generation as a between-subjects factor and experimental phase (learning vs. communication) as a within-subjects factor found a significant linear trend for Generation, $F(1, 35) = 17.49$, $p < .001$. d_{unb} for the increase over generations post-learning was 1.47, and post-communication was 1.54, suggesting a very large effect. There was also a significant main effect of experimental phase, $F(1, 35) = 30.70$, $p < .001$. The interaction was not significant, $F(4, 35) = 0.30$, $p = .88$, showing that the linear trend was similar for the two phases. d_{unb} values for the increase in convexity from the end of learning to the end of communication ranged from 0.56 in generation 3 to 1.35 in generation 2, suggesting a consistent medium-to-large effect of communication in increasing the convexity of participants' category systems.

A one-sample t-test confirmed that the mean convexity of the generation 1 category systems at the end of communication was not significantly different from the convexity of the similarity baseline category system (0.51), $t(7) = -0.73$, $p = .49$. By generation 5, however, the mean convexity of post-communication category systems was significantly higher than the baseline, $t(7) = 4.65$, $p = .002$. For systems after the end of learning, conversely, their convexity started as significantly lower than baseline in generation 1, $t(7) = -3.82$, $p = .007$, and by generation 5 was not significantly higher than baseline, $t(7) = 0.48$, $p = .65$. Therefore, at least in the generational time of this experiment, learning alone does not lead to category systems whose convexity is above the similarity baseline.

5.4.2.5 Category system alignment

Alignment of pairs' category systems at the end of communication increased over generations (Figure 5.15). In generation 1 alignment at the end of communication was 0.21 [0.09, 0.33]. A t-test found that this was statistically comparable to the value of 0.24 [0.08, 0.40] for the communicators from Experiment 3, $t(16) = -0.29$, $p = .77$, despite the Experiment 4 pairs having learned the same system first. Alignment increased by generation 5 to 0.53 [0.36, 0.70]. The increase from generation 1 to generation 5 was 0.32 [0.15, 0.49]. By the end of communication in generation 5, the average alignment value of 0.53 was statistically comparable to the value of 0.48 from whole-set

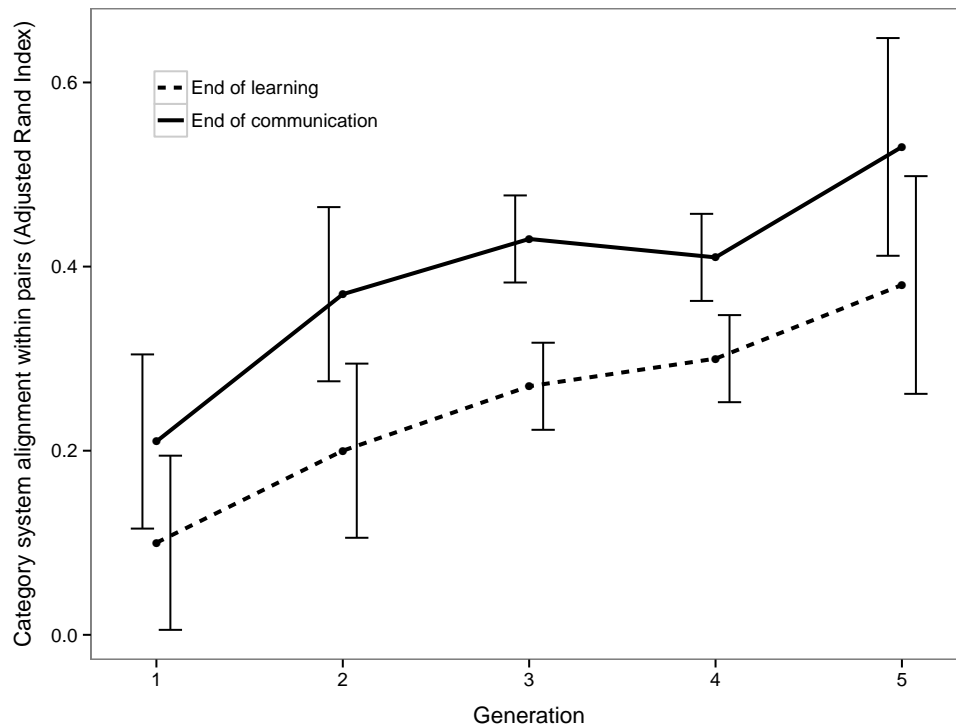


Figure 5.15: Graph showing change in pairs' category system alignment over generations in Experiment 4. Dotted line shows alignment post-learning, solid line alignment post-communication. Error bars show 95% confidence intervals, with standard errors adjusted to reflect only between-subjects differences.

non-communicators in Experiment 3, $t(16) = 0.57$, $p = .57$.

We can also compare how aligned pairs' category systems were at the end of the learning phase, to determine how much of their alignment was already in place as a result of having learned the same system. Alignment at the end of the learning phase was consistently lower than at the end of the communication phase. Post-learning alignment was 0.1 [0.01, 0.19] in generation 1, rising to 0.38 [0.14, 0.62] by generation 5. This was an average increase of 0.28 [0.03, 0.53].

Across all generations, the mean difference in alignment between the end of learning and the end of communication was remarkably consistent. In Generation 1, pairs' alignment increased by an average of 0.11 [-0.08, 0.30] between the end of learning and the end of communication; in Generation 2, the increase was 0.17 [-0.02, 0.36]; in Generation 3, 0.16 [0.04, 0.28]; in Generation 4, 0.11 [-0.01, 0.23]; and in Generation 5, 0.16 [-0.05, 0.37].

A mixed ANOVA with Generation as a between-subjects factor and experiment phase (learning vs. communication) as a within-subjects factor found a significant lin-

ear trend for Generation, $F(1, 35) = 9.70$, $p = .004$. d_{umb} for the increase in alignment over generations post-learning was 1.19, and post-communication was 1.75, suggesting a very large effect. There was also a significant main effect of experiment phase, $F(1, 35) = 18.79$, $p < .001$. The interaction was not significant, $F(4, 35) = 0.14$, $p = .97$, showing that the linear trend was similar for the two phases. d_{umb} values for the increase in alignment between learning and communication ranged from 0.5 (Generation 4) to 0.85 (Generation 2), suggesting a consistent medium-to-large effect of communication in increasing alignment. Taken together with the lack of interaction, the confidence intervals and d values for the differences between learning and communication in each generation above suggest that the increase in alignment between the end of learning and the end of communication is fairly consistent across generations.

As in the analysis of Experiment 3, Monte Carlo simulations were run to determine whether category systems at the end of communication were more aligned within than across pairs. For generation 1, the probability of across-pair alignment being higher than within-pair alignment was .03; in generation 2, .006; in generation 3, .0006; in generation 4, .004; and in generation 5, .004. Thus, alignment within pairs was reliably higher than across pairs.

Figure 5.16 shows average alignment within pairs, with the mean value for alignment across pairs as a comparison, for all conditions in Experiment 3 and all generations in Experiment 4, measured for category systems at the end of communication.

As Figure 5.16 shows, alignment across pairs (i.e. across different chains, between participants who never interacted) increased over generations, reaching 0.43 by generation 5. These category systems were therefore 43% more aligned than would be expected by chance, suggesting some partial convergence to similar category structures across chains. However, the higher level of alignment within pairs shows that lineage-specific variation remained.

One possibility is that over generations, participants' category systems were converging by more closely reflecting a similarity-based categorisation. We can test this by looking at levels of alignment with the similarity baseline category system. These results are shown in Figure 5.17.

At the end of the learning phase in generation 1, alignment with the baseline system was 0.08 [0.03, 0.13]; this rose to 0.22 [0.15, 0.29] by generation 5. The overall increase from generation 1 to 5 at the end of learning was 0.14 [0.06, 0.22]. At the end of the communication phase in generation 1, alignment with the baseline system was 0.20 [0.15, 0.25]; this rose to 0.31 [0.26, 0.36] by generation 5. The overall increase

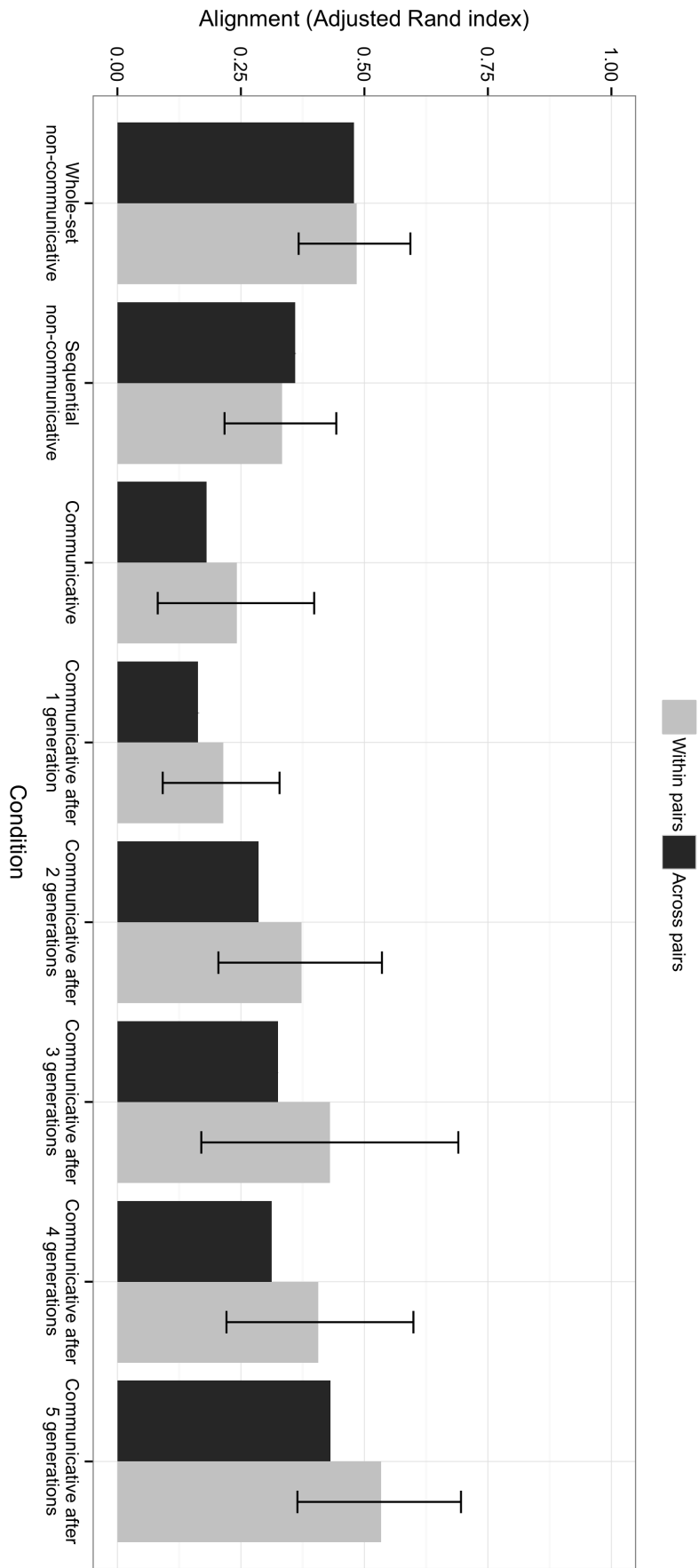


Figure 5.16: Graph showing alignment (adjusted Rand index) within pairs (light bars) and across pairs (dark bars) for output systems from Experiment 3, and output systems from the end of the communication phase at each generation in Experiment 4. Error bars are 95% confidence intervals. Across-pair mean alignment values are generated from simulation and hence do not have confidence intervals; see text for statistical comparison of mean within-pair alignment values to the simulated across-pair distribution.

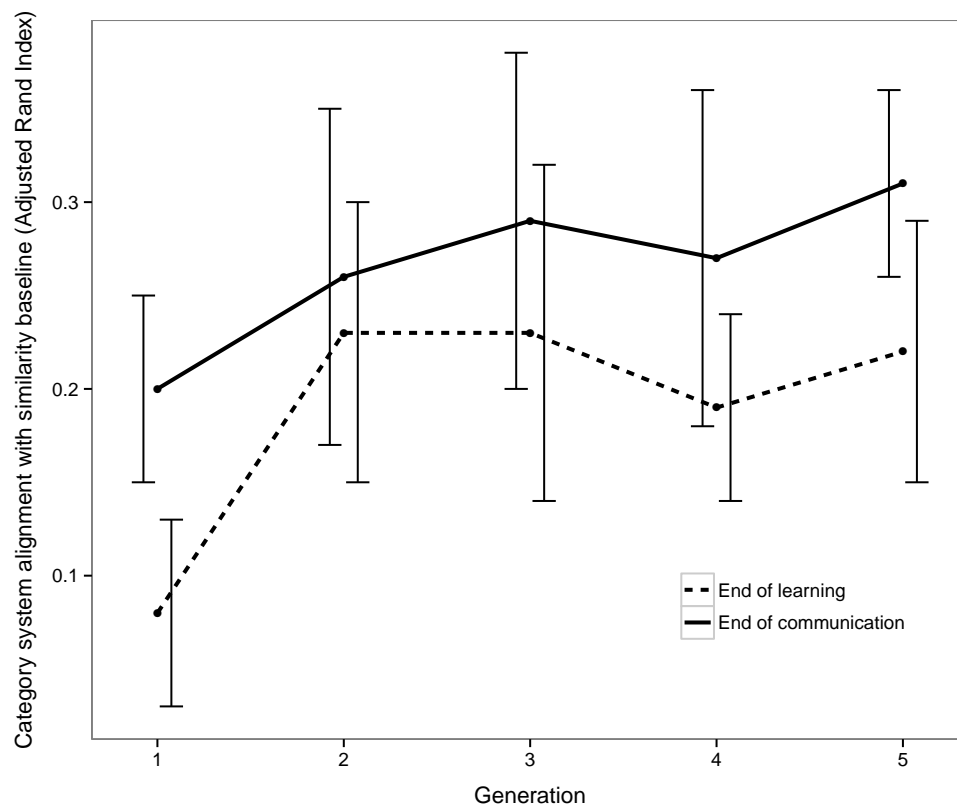


Figure 5.17: Graph showing alignment (adjusted Rand index) with the baseline system generated from participants' pairwise similarity ratings in Experiment 2. Error bars are 95% confidence intervals.

from generation 1 to 5 at the end of communication was 0.11 [0.04, 0.22]. Alignment with the baseline system generally increased between learning and communication: these increases were, in generation 1, 0.12 [0.05, 0.19]; in generation 2, 0.03 [-0.06, 0.12]; in generation 3, 0.06 [-0.03, 0.15]; in generation 4, 0.08 [0.01, 0.15]; and in generation 5, 0.09 [0.02, 0.16].

A mixed ANOVA, with Generation as a between-subjects factor and experiment phase (learning vs. communication) as a within-subjects factor, found a significant linear trend for Generation, $F(1,35) = 8.16$, $p = .007$. d_{unb} values for the increase in alignment with the baseline system over generations were 1.82 for post-learning, and 1.60 for post-communication, suggesting a very large effect. There was also a significant main effect of experiment phase, $F(1,35) = 27.02$, $p < .001$. There was no significant interaction, $F(4,35) = 1.02$, $p = .41$. d_{unb} values for the increase in alignment with the baseline between learning and communication were above 0.5 except for in generation 2, where d_{unb} was 0.28. This suggests a relatively consistent medium-to-large effect of communication in increasing alignment to the baseline system. Taken together with the confidence intervals above, this suggests that communication generally encouraged higher convergence to similarity-based structures than learning alone, although this result is less clear in generations 2 and 3. The highest alignment reached with the similarity baseline was 0.31, showing that the increasing amounts of across-pair alignment over generations cannot be explained only in terms of convergence towards similarity-based structures.

5.4.3 Experiment 4 Discussion

Despite the change of having an initial learning phase, the key results from Experiment 3 are replicated in Experiment 4: communicative success, category system convexity, and alignment after communication in generation 1 were all statistically comparable to the communicative condition from Experiment 3.

Category systems changed over generations to become more learnable, as measured by participants' accuracy both in applying words to images, and in reproducing the structure of the category system they were trained on. At least two factors potentially contribute to this change: 1) category systems lose distinctions over generations, meaning that there are fewer words to learn; 2) category systems become more convex over generations, rationalising the groups of images that particular words refer to. While the loss of distinctions appears to be a consequence of iterated learning,

the increase in convexity appears to be mostly due to communication, since convexity of category systems drops during each learning phase. This throws some doubt on the idea that more convex categories are necessarily better for learning. Indeed, if participants are using similarity-based generalisation strategies to acquire the category systems during learning, this may lead them to infer less convex categories than those they are presented with: see the relatively low convexity score for the similarity baseline category system. However, it could simply be the case that the time constraints of the learning phase are not sufficient for participants to fully learn the category system they are trained on, leading to the drop in convexity during this phase. This possibility is considered further below in section 5.6.

Participants in generation 1 were trained on a system with 25 categories. Accordingly, they used more categories than their counterparts in the communicative condition of Experiment 3. In subsequent generations, the differing pressures of learning and communication become clear: categories tend to be lost during learning, and then regained during communication, until the number of categories stabilises at around 7.5 in generation 5. This was close to, but statistically higher than, the number produced by non-communicating participants in Experiment 3, speaking to the increased incentive for expressivity in communication. This level of expressivity was more consistent across pairs than the expressivity of communicators' systems in Experiment 3.

Despite this drop in number of categories, communicative success increased reliably over generations, showing that pairs structure and align their category systems to compensate for the loss in precision. As noted above, this increase in convex structure and alignment appears to happen within the communication phase, rather than the learning phase of the experiment. Both category convexity and alignment dropped during learning and increased during communication, a pattern which held even in later generations, when error was lower and systems were generally being more faithfully acquired, showing this was not just an effect of poor learning being rectified during communication. This shows that communication works as hypothesised in the Introduction to structure and align category systems; however, for these pressures to take effect, the common ground of a mutually acquired category system with a learnable number of distinctions must be in place. A learnable category system produced via cultural transmission, with communication working on top of the common ground this learning provides, leads to aligned, convex categories.

Do the pressures of learning and communication push systems to converge towards universal structures, or (as discussed in section 3.5.3 of Chapter 3) are category sys-

tems constrained into particular paths by the inertia of established conventions? The answer appears to be a little of both. Participants who did not interact converged over generations onto somewhat similar category systems (as evidenced by higher levels of alignment across pairs in later generations). The hypothesis that the systems are converging closer to a pre-existing similarity space was partly supported by the increase in alignment with the similarity baseline system over generations. However, this alignment is not high enough to explain the full extent of the convergence. This suggests that the system that participants are converging towards is defined not only by similarity but also by the pressures imposed by the experimental task. One possibility is suggested by the reliable increase in category system convexity over communication at each generation. If communication creates a pressure towards more convex categories, this could lead to convergence across chains (since convexity constrains the possible shapes of categories, leading to more alignment as a by-product).

Importantly, however, convergence is not total: the resulting systems still vary according to lineage-specific conventions established by particular pairs in a chain's history (as evidenced by higher levels of alignment within than across pairs over all generations). Figure 5.18 shows the progression of a typical chain in Experiment 4, illustrating the changes in category structure that occur over generations of learning and communication.

5.5 Exploratory results

Beyond the dependent variables outlined in section 5.2.2, the results also offer avenues for more speculative analysis. Considered below are factors associated with communicative success, and areas of particular consensus on category boundaries in the stimulus space.

5.5.1 Factors affecting communicative success

5.5.1.1 Experiment 3

As noted in the main Results above, in the communicative condition of Experiment 3, there was no significant correlation between average number of categories a pair used and communicative success, $r = -0.30$ $[-0.78, 0.41]$, $p = .41$. Category system convexity, on the other hand, was significantly correlated with communicative success in the last two rounds, $r = .81$ $[.38, .95]$, $p = .004$, as was alignment, $r = .72$ $[.17, .93]$,

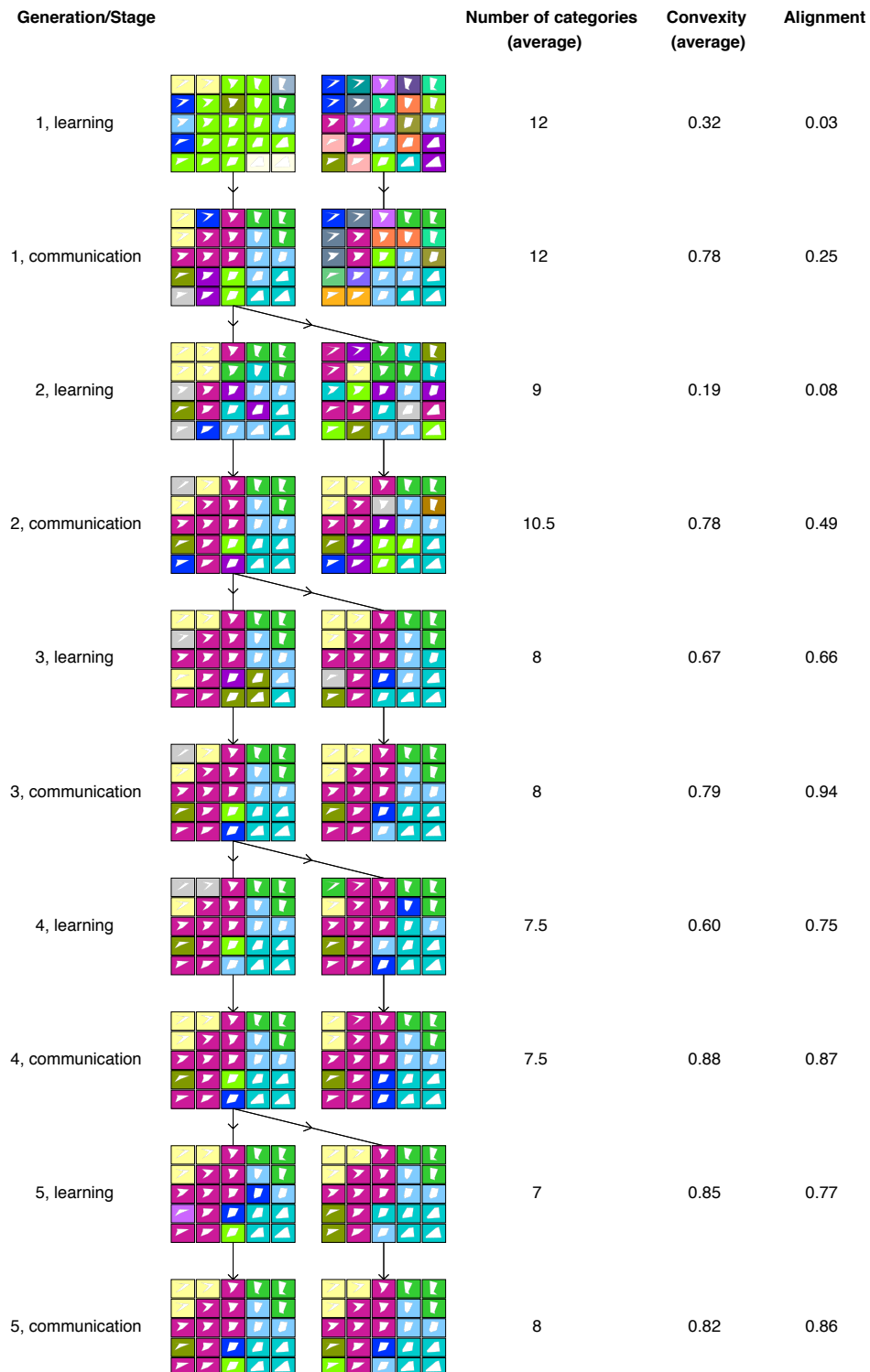


Figure 5.18: Progression of chain 2 in Experiment 4, showing category systems at the end of learning and communication for each generation. The system passed on to the next generation as learning input is always shown on the left.

$p = .02$. These correlations suggest that rather than the number of distinctions available in a system, it is the structure of categories and their alignment between interlocutors that is crucial to success in the communication game.

Going beyond the labelling behaviour of senders, an intriguing question is the extent to which receiver behaviour contributes to communicative success. For example, a receiver could choose always to pick the same image for a given word. If their partner also held to this same behaviour, and the image was highly salient to them both, this could result in a relatively high level of success, even if their labelling behaviours were not completely aligned. There is some evidence that pairs who coordinated successfully on this ‘preferred exemplar’ strategy had higher communicative success. To investigate this, preferred exemplars were defined as the images the participant selected with the highest frequency when their partner used a given word. If the same image was among the preferred exemplars for both participants in a pair for the same word, they were counted as sharing a preferred exemplar for that word. The proportion of words for which they shared preferred exemplars was then calculated. Using this measure, communicating pairs in Experiment 3 who had a higher proportion of shared preferred exemplars for particular categories tended to be more successful, $r = .85$ [.48, .96], $p = .002$.

Participants often had more than one preferred exemplar for a given word; unambiguously preferred exemplars were too rare to allow a systematic analysis of whether they tended to be located at the centre of categories. A qualitative analysis found that while some are central, there is a tendency for corner or side images to be preferred exemplars (perhaps because, given the nature of the morphed space, they are more distinctive from other images and hence more salient). Figure 5.19 (left panel) shows an example of one receiver’s preferred exemplars, superimposed on their partner’s labelled categories. While there is some evidence of central preferred exemplars (e.g. turquoise category), most are edge or corner images, even where these are not central to their respective categories (e.g., blue and magenta categories).

A more general question concerns the internal structure of participants’ categories. Do they treat the images in a given category as exemplars, or do their categories have a more prototype-like structure? This question is difficult to answer by looking at these data. Modelling approaches offer a fruitful avenue for investigating this further. Collaborative work with Bill Thompson found that participants’ categorisation behaviour across the three conditions of Experiment 3 can be captured by a hierarchical Bayesian model with two key parameters: α , which quantifies the extent to which the internal

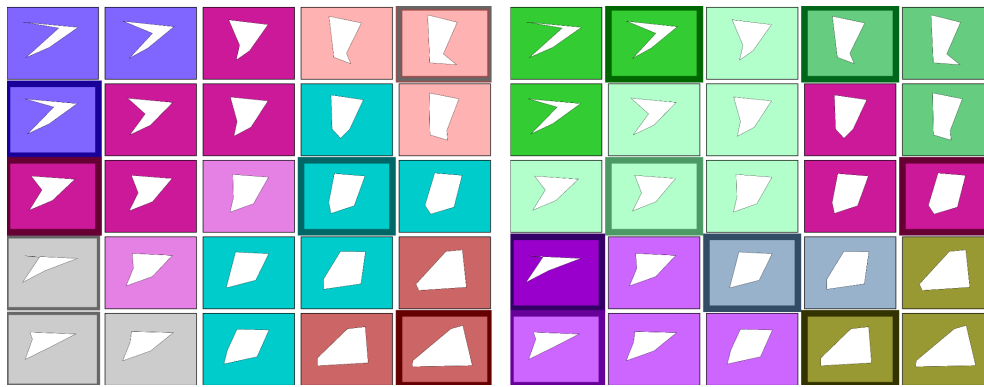


Figure 5.19: Two examples of receivers' preferred exemplars, taken from the communicative condition of Experiment 3 (left panel) and from the end of communication in generation 5 of Experiment 4 (right panel). Bold dark-coloured lines indicate the preferred exemplar for the sender's corresponding light-coloured category. Thinner lines indicate where two or more images were equally preferred. Most preferred exemplars are from the sides or corners of the stimulus space, perhaps showing a preference for more distinctive images.

structure of participants' individual categories is exemplar-like (consisting of many small clusters) or prototype-like (consisting of one cluster), and β , which quantifies how willing participants are to create new categories, rather than accepting new images into existing categories. The model found that while all participants' categories were fairly exemplar-like, communicators' categories were more prototype-like than those of non-communicators; communicators were also more conservative about accepting new stimuli into existing categories, preferring to create new categories (Thompson et al., 2014). This tentatively suggests that communication incentivises more categories with a more unified internal structure (i.e. clustering around a single prototype), while individual categorisation may be more tolerant of heterogeneous categories composed of several clusters. Future work extending this model could provide further insight into the internal structure of participants' categories.

5.5.1.2 Experiment 4

To investigate whether any of the above factors were significant predictors of communicative success in Experiment 4, a regression analysis was run incorporating the following as predictors: average number of categories, category system convexity, alignment, and proportion of categories where participants had shared preferred exemplars. The overall model fit was $R^2 = .54$, explaining around half the variance in commu-

nicative success in the last two rounds. However, alignment was the only significant predictor, $\beta = 89.8$, $p = .006$. This result is interesting in the light of Gärdenfors and Warglien's claims that convexity is crucial to communication, with alignment being optional: 'semantic equilibria can exist without needing to assume that the communicating individuals possess the same mental spaces...the shapes of our conceptual structures make it possible to find a point of convergence' (Warglien & Gärdenfors, 2011, p. 2167-9). However, it is important to bear in mind that alignment and convexity of categories are highly correlated, $r = .73$ [.54, .85], $p < .001$, which may lead to convexity not emerging as an independently significant predictor. There are less strong but still significant correlations between proportion of preferred exemplars shared and both alignment ($r = .39$ [.09, .63], $p = .01$) and convexity ($r = .32$ [.01, .58], $p = .04$). These correlations are in themselves interesting. Possibly, convex categories are easier to align on, or the mechanisms by which communicators align naturally lead to convex categories. Convex categories may build up around preferred exemplars, as Gärdenfors suggests (although see Figure 5.19 for evidence that the preferred exemplars of these categories are not necessarily central to the category, as in Gärdenfors's model). Alignment on labelling and reception behaviour may proceed in parallel, as per the interactive alignment account (Garrod & Pickering, 2009).

In interpreting these results, it is important to bear in mind that these results may vary substantially over different generations. To check this, generation was incorporated into another regression model, and did not affect the results – alignment was still the only significant predictor, $\beta = 93.22$, $p = .005$. However, the sparsity of data points ($N = 8$ for each generation) means these hypotheses about contributors to communicative success must remain speculative.

5.5.2 Consensus on category boundaries

The higher levels of alignment in the whole-set non-communicative condition in Experiment 3, and across pairs in later generations of communication in Experiment 4, suggest greater consensus across the whole group of participants in their category systems. Is this consensus spread evenly across the set of images, or does agreement cluster around particular groups of images? Figure 5.20 shows graphically the level of consensus regarding category boundaries in the conditions of Experiment 3 and the generations of Experiment 4 (post-communication). Here, we can confirm that consensus is highest in the whole-set non-communicative condition of Experiment 3

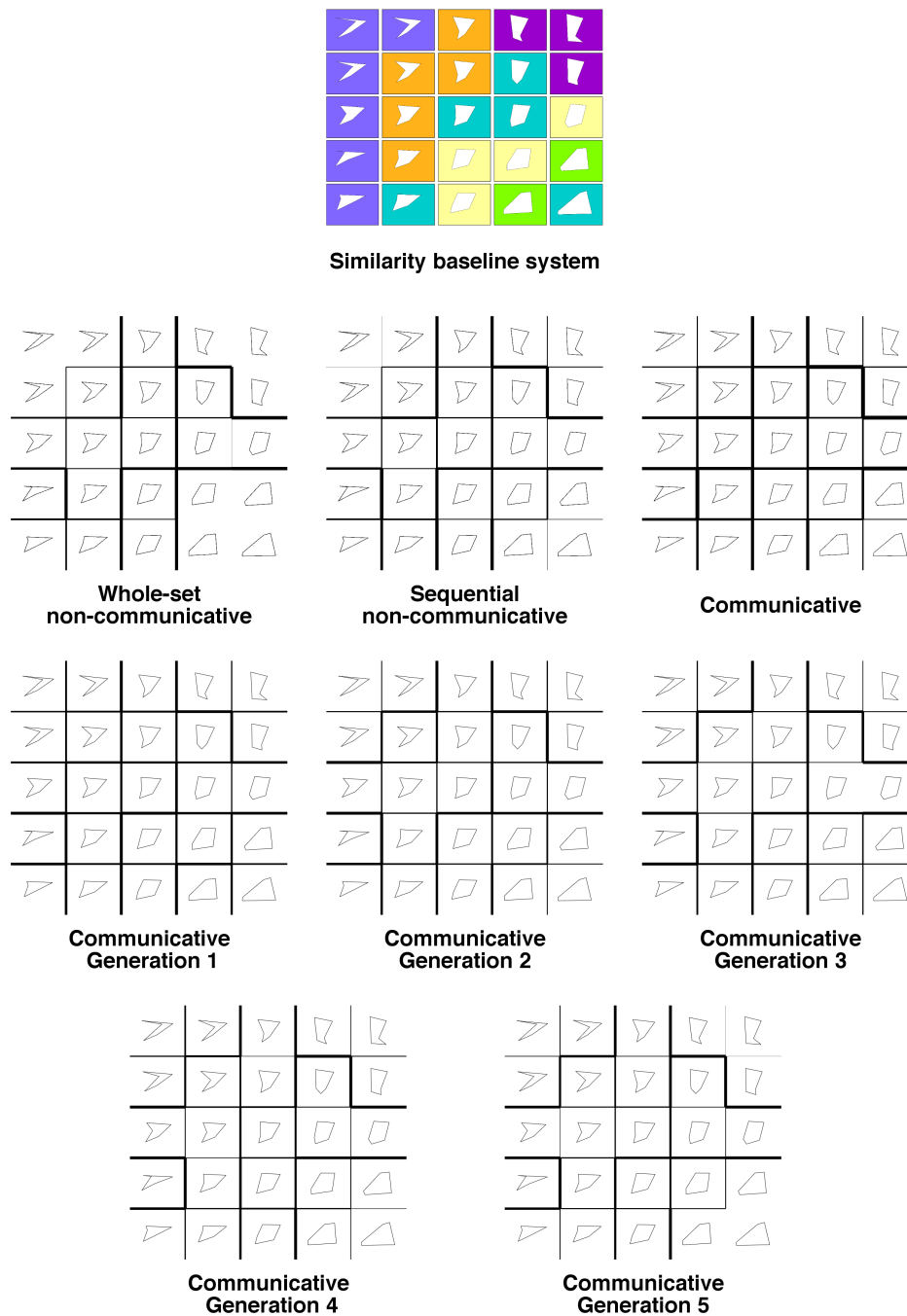


Figure 5.20: Graphical representation of the level of consensus between participants on the placement of category boundaries in the conditions of Experiment 3 (second row) and the generations of Experiment 4, after communication (third and fourth rows). Thicker lines mean more participants agreed on a boundary; thinner or absent lines mean more participants agreed on the absence of a boundary. Medium-weight lines mean a lack of clear consensus.

and in generation 5 of Experiment 4; however, this consensus is focused on particular groups (regions with absent or thicker lines, mostly the corners) with other parts of the space less clearly discretised (regions with medium-weight lines, mostly the middle). Comparing this to the baseline similarity category system, we can see that while some regions of consensus correspond to regions in the baseline system (for example, the three images in the top right corner), others do not correspond so clearly (for example, the bottom right and top left corners). As previously suggested, the task of categorising images may impose different structure on the space than can be derived from pairwise similarity.

5.6 Design issues and future directions

There are a number of areas where the design of the experiment could be improved. As noted in section 5.2.2.2, the convexity measure, as well as the success scores in the communication game, assume that the Euclidean distance metric used to construct the stimuli is the relevant similarity for evaluating participants' behaviour. This assumption may not hold. With regard to the convexity assumption, the relative lack of convexity of the category system produced from perceptual similarity ratings (Figure 5.5) suggests that pairwise perceptual similarity in the stimulus space does not straightforwardly reflect inverse Euclidean distance: if this were the case, similarity ratings clustered via the k -means algorithm would produce convex categories. With regard to communicative success, participants' questionnaire responses revealed that some scores did not correspond well with the apparent perceptual similarity of the target and the selected image. That participants were at least partly using non-Euclidean dimensions to classify the images is also apparent from questionnaire responses, where some participants reported classifying images by number of sides or angularity. Future work could investigate the extent to which these self-reports of categorisation rules correlate with the category structures produced during the course of the experiments.

The stimulus space in the experiments was designed to minimise intrinsic structure or obvious category boundaries. This kind of continuous space for categorisation does not have many obvious analogues in the real world: even in the technically continuous space of colour, some points are more salient or focal than others (e.g. Regier et al., 2007). However, the intention in this experiment was to model a space without rich pre-existing structure in order to be able to see more clearly what effect, if any, learning and communication had on the structure of categories in this space. Despite

this effort, the uneven similarity gradient caused by the edges and corners (with these images being necessarily more distinct from the rest of the set on average than images in the middle) means that some images are more salient than others. This may lead participants to pick these as preferred category exemplars (Figure 5.19) and possibly constrain the development of category systems. However, there were still detectable differences between the conditions in the category structures produced, showing that the perceptual structure of the stimulus space still allowed for variation. Future work could consider how different kinds of stimulus spaces might interact with learning and communication pressures, and which meaning spaces are better models of the real world. A related issue is that while the similarity baseline system created from pairwise ratings is used for comparison throughout, the k -means procedure that was used to create it assumes categories with central prototypes, whereas the participants' reception behaviour suggests that their categories did not necessarily follow this constraint. Alternative methods of obtaining a baseline category system, for example by using different clustering algorithms, could provide a more comparable result.

The validity of the conditions of Experiment 3 as a comparison of individual categorisation to communication can also be debated. The whole-set non-communicative, sequential non-communicative and communicative conditions exert very different task demands, which may not be reducible to the difference between individual categorisation and communication. Modelling one experimental proxy for individual categorisation is in itself challenging: in the real world, we do not categorise objects purely according to perceptual similarity. Future work could provide a more task-based, functionally relevant task to elicit categorisation, such as grouping unfamiliar creatures based on whether they should be approached or avoided/captured or destroyed (as in Lupyan et al., 2007; Voiklis & Corter, 2012). The latter study, which contrasts the effectiveness of individuals at learning these categories with that of pairs engaging in dialogue, could be adapted into an artificial language learning paradigm without pre-existing learned categories, potentially providing a more balanced comparison of the category structures that emerge in each condition.

Furthermore, the task in the communication game – to identify a unique exemplar from the whole set of images – is both a very different task from the categorisation task performed in the individual conditions, and not necessarily an ecologically valid proxy for communication in real language. If I tell you that I am sitting on a chair, it does not necessarily matter if you are picturing an armchair where I meant a dining chair, unless the particular characteristics of the chair become relevant for the story I am telling. In

the experiments, the specificity of the feedback in the communication game (where the target exemplar is provided and the score is based on the distance of the selected image from this target) may encourage artificially fine-grained categories compared to real language, where communication can proceed unharmed as long as our words refer to the same broad categories of objects. Future work could use different feedback and scoring mechanisms to provide a more ecologically valid communicative setup: for example, instead of a graded score proportional to Euclidean distance of the target from the selected image, participants could be given a ‘correct’ response if their selected image was within a certain distance of the target, and an ‘incorrect’ response otherwise, without being informed of the specific target image.⁹ This could potentially encourage more convex categories, but could also actually hold back alignment – participants could have higher levels of disagreement about the particular groups of exemplars associated with words while still having high levels of communicative success. In summary, given the limitations of the model of communication used in the experiments, other communication paradigms should be explored to ascertain the generality of the results. These issues will be discussed more generally in Chapter 7.

Another potential concern, alluded to in the Discussion of Experiment 4, is that the increases in convexity and alignment during the communication phase simply reflect participants’ greater experience with the stimuli, rather than being specifically a result of communication. Perhaps participants do not have enough time in the learning phase to fully acquire the category system they are being trained on, and only complete this learning process during the communication phase (and in doing so, align more closely). However, as noted above, the fact these increases are consistent in size over generations, even though error is dropping and the systems are becoming easier to learn, suggests that this is a genuine effect of communication rather than an artefact of having extra learning time. To substantiate this, however, an ideal solution would be to run a control experiment. Instead of iterated chains of learning and communication, the control experiment would feature only iterated learning of category systems, with an overall equivalent amount of exposure to the stimuli as in the learning and communication phases of Experiment 4. The result of this experiment would tell us whether the increase in convexity and alignment at each generation is specifically due to communicative pressures or is just a result of improved learning.

The results concerning the factors that are most important for communicative success are for the moment speculative. An interesting extension would be to design

⁹Thanks to Holly Branigan for this idea.

category systems with particular characteristics, train groups of participants on these systems, and then compare their success scores in the communication game. Preliminary simulations, not reported here, have found that, given certain assumptions about communication strategies, more convex category systems and higher alignment within pairs do lead to higher communicative success. It would be interesting to extend and substantiate these results in human populations.

5.7 Conclusion

This chapter began with the observation that individual categorisation and communicative word reuse are two ways of dividing a continuous world up into discrete chunks. The first question we set out to answer was: do individual categorisation and communication encourage different kinds of category structure? More specifically, does communication encourage systems that are more expressive, more convex, and more aligned within pairs? The second question was: how does cultural transmission interact with communication to shape word meaning structures?

The results were somewhat surprising. While communication alone produces more expressive category structures than individual categorisation, the level of expressivity varies widely across pairs. In addition, the category structures produced by communication alone are less convex and less aligned than those produced by individual categorisation. Given that higher convexity and higher alignment is associated with greater communicative success, this suggests that non-communicating individuals in Experiment 3 actually produced category systems which were more optimal for communication than those of their communicative counterparts. Both sets of non-communicators produced category structures that were more convex than those produced by pairwise similarity ratings, supporting Gärdenfors's position that the convexity of concepts is a property of individual cognition that does not necessarily require convex structure in the world. However, the low levels of convexity in communicators seem to speak against his position that convexity is also optimal for communication.

A possible explanation for this puzzling result is the unique set of constraints on communicators in this task. Not only does the memory burden of labelling images sequentially disrupt category convexity and alignment (as seen from the lower results on both measures in the sequential non-communicative condition), but the additional pressure of having to coordinate with a partner encourages agreement on conventions for particular images, rather than alignment on optimal category systems. This result is

at odds with previous work showing that communication works to optimise and align interlocutors' categories. The key difference may be that in the real world, communication takes place on the back of a learned linguistic system that acts as common ground (see section 3.5.3 in Chapter 3).

Experiment 4 followed up on this observation to test how cultural transmission of a learned system interacts with the effect of communication. Repeated cycles of iterated learning and communication worked to produce category systems that were more faithfully reproduced, more stable in number of categories, more convex, and more aligned within pairs over generations. While the number of categories in the systems dropped over generations, it remained consistently higher than that of systems produced by individual categorisation. In addition, while the number of categories dropped, communicative success increased, showing that categories were becoming more optimised for communication. By the end of generation 5, levels of convexity and alignment in communicators' category system were comparable with the whole-set non-communicators in Experiment 3.

Many of these features of category systems may be useful for individual learning as well as for communication. For example, convex categories may be advantageous for generalisation in learning; alignment may result from similar learning capacities being applied to similar environmental input, rather than a specific adaptation for communication. The results of Experiment 3 show that high levels of convexity and alignment can result from shared individual biases where categorisation decisions are made on a simultaneous basis. However, encountering items one by one disrupts this process, and the dual task of having to learn a partner's system simultaneously with creating one's own disrupts it further. The results of Experiment 4 show that to achieve comparably high levels of convexity and alignment under these conditions, a combination of iterated learning (to cut down distinctions to a learnable number) and communication (to structure and align category systems on the basis of this learned common ground) are required. In contrast with categories produced according to the shared biases of individuals, categories produced over iterated learning and communication show higher alignment within the pairs that produced them than across the whole population. This difference shows how the process of establishing communicative conventions leads to both universal aspects of communicative category systems (partly reflecting shared task pressures and partly converging towards perceptual similarity), and more idiosyncratic features contingent on a pair's particular interaction history. This provides an experimental result that parallels the mixture of convergence and diversity in commu-

nicative categories we see across natural languages (Majid et al., 2008), and illuminates the unique contributions of iterated learning and communication to the structure of word meanings.

The interaction of word meanings with event structure: Words are generalised on the basis of unpredictable event features

6.1 Introduction

A basic function of words is labelling objects and actions (Heine & Kuteva, 2002; Deutscher, 2005). However, objects and actions do not present themselves to be labelled in isolation. Instead, we encounter them as parts of complex events in the world. The object and action combinations we see are constrained, by everything from physics to stereotypes. If I drop a rock, it does not drift gently downwards; grandmothers are, on average, unlikely to practise capoeira. These constraints, be they absolute or probabilistic, are part of our general world knowledge.

The pattern of likely events in the world is dependent on a number of factors. Firstly, different objects have different levels of constraint concerning the actions they can participate in. For example, while both rocks and kittens can be picked up and dropped, kittens can take part in a more varied range of events than rocks can (eating, purring, being stroked, etc.). Secondly, sequences of actions are conditioned on each other and on the objects that are involved. If I throw a rock into a swimming pool, it

will sink to the bottom. If I throw a kitten into a swimming pool, the results are more unpredictable.

Knowledge of these object-action dependencies is a key part of our language production and comprehension. Language users continually integrate their knowledge of conventions in the language with their knowledge of events in the world to predict utterance meaning (for reviews see Hagoort & van Berkum, 2007; Altmann & Mirković, 2009). For example, listeners look towards edible objects on hearing the word ‘eat’ (Altmann & Kamide, 1999); exposure to nouns primes other nouns associated with typical events they play a role in, e.g., ‘sale’ primes ‘shopper’ (Hare et al., 2009); and readers are faster to process patients when they follow agent-action combinations (Bicknell et al., 2010) or instrument-action pairs (Matsuki et al., 2011) that strongly predict them. For example, ‘The journalist checked’ facilitates processing of ‘spelling’, whereas ‘The mechanic checked’ facilitates ‘brakes’; ‘used the shampoo to wash’ facilitates ‘hair’, whereas ‘used the hose to wash’ facilitates ‘car’. Importantly, these effects are not due to simple word association, or to direct agent-patient or instrument-patient relations, but require knowledge of complex patterns of event typicality. These constraints are strong enough even to compensate for other sources of processing difficulty, such as ‘garden path’ effects stemming from syntactic ambiguity. McRae & Matsuki (2013) point out that sentence 2, structurally identical to sentence 1, does not induce the same ‘garden path’ effect, due in part to the fact that landmines are typical patients of burying and very atypical agents:

- 1) The horse raced past the barn fell.
- 2) The landmine buried in the sand exploded.

The predictive value of event knowledge is interdependent with the patterns of reuse of words in a particular language. These conventions determine which event features are more strongly associated with the word, and hence in a sense provided, and which must be inferred from context. Recalling John Searle’s example from Chapter 2, ‘cut the grass’ and ‘cut the cake’ use the same verb to describe two very different events. The events are similar in that they both involve making a separation using a sharp tool; however, perceptually and functionally, they involve very different instruments, methods, and results. However, when we describe these events, we do not have to specify that the cutting event in one case involves a lawnmower and results in shorter grass, and in the other case involves a knife and results in edible-sized pieces of cake. These event features are co-redundant with the objects involved, and hence do not need

to be separately expressed (and indeed, if they are, the effect is extremely odd: ‘I cut the grass using a lawnmower to make it shorter’.) If we use ‘cut’ and ‘grass’ together, we can infer the details from what we know about grass and the kind of things that are usually done to it. The conventions in a language and the patterns of event structure in the world therefore work together in a complementary fashion to constrain the interpretation of utterances.

This has implications for the hypothesis that communication optimises word meanings to be expressive. As noted in Chapter 3, expressivity does not entail lexicalising every perceptible feature in the world: it involves maximising the likelihood of inferring utterance meanings across communicative contexts. If words evolve to function as efficient clues for inferring utterance meaning in interaction with world knowledge, we can make an evolutionary prediction: the event features lexicalised in a language will tend to be those that are less predictable on average across all utterances, given the structure of events in the world and the conventions previously established in the language. This account offers an explanation for both universal tendencies in word meanings, given the relative invariance of event structure across language communities, and language-specific variation, given the diverse and continually changing conventions of different languages.

Experiments 1 to 4, presented in Chapters 4 and 5, have shown that learning and communication act to shape word meanings in interaction with the structure of the world; however, the meanings involved were noun-like and labelled in isolation. Experiment 5, presented in this chapter, shows how these pressures interact when the world is structured not only by the perceptible features of objects, but also by co-occurrence patterns of objects and actions in complex events.

6.2 Background

As reviewed above, eyetracking and fMRI studies have shown that listeners integrate their knowledge of how actions and objects typically combine at the earliest stages of language comprehension. For example, when participants hear ‘The boy will eat’, their eye movements are already preferentially directed towards referents that are plausible patients for this action, such as a picture of a cake (Altmann & Kamide, 1999). Pirog Reville et al. (2008) showed that these incremental prediction effects can also be induced for artificial languages that refer to events in an artificial world. Participants were trained on an artificial lexicon with words referring to objects and actions. They

were also trained on events involving these objects and actions, where particular objects were constrained to perform particular actions: for example, only curved objects moved, while only straight objects changed colour. Subsequently, when they heard utterances in the artificial language, participants showed the same kind of predictive behaviour as participants in natural language studies. On hearing an action label, they anticipatorily looked at the object which was constrained to perform this action (Pirog Revill et al., 2008, p. 1219).

The experiment presented in this chapter uses stimuli adapted from Pirog Revill et al. (2008). However, in the current study, rather than being trained on a language with words for each event feature, participants are trained on two labels which refer to complex events. They then have to generalise this language over the course of communication with their partner. Participants therefore have to make and continually update their inferences about which event features should be lexicalised. Participants' inferences are tracked firstly via their patterns of word reuse during communication, and secondly via a post-test where they are asked to label complex events and the objects and actions that compose them. The use of a structured world recalls some of the modelling work discussed in Chapter 3; however, rather than using noun-like meanings in isolation, the stimuli in Experiment 5 feature complex event structure patterns, where particular objects and actions co-occur in non-random ways.

The aim is to investigate the effect of these event structure constraints on how participants generalise and change a small learned lexicon during and after communication. Where Experiment 1 investigated the interaction of world structure and iterated learning, and Experiments 2, 3 and 4 investigated the interaction of learning and communication where world structure is less discrete, Experiment 5 investigates the interaction of world structure, learning, and communication: specifically, how knowledge of typical events in the world interacts with the reuse and generalisation of learned conventions to influence the event features that are lexicalised in a novel artificial language.

6.3 Methods

The experiment consisted of five phases. In the first phase, participants were shown a series of possible events in an 'alien' world. They were instructed to watch the events and try to get an idea of which events were possible and impossible in this world. The idea was to familiarise participants with typical events, such that they would build up

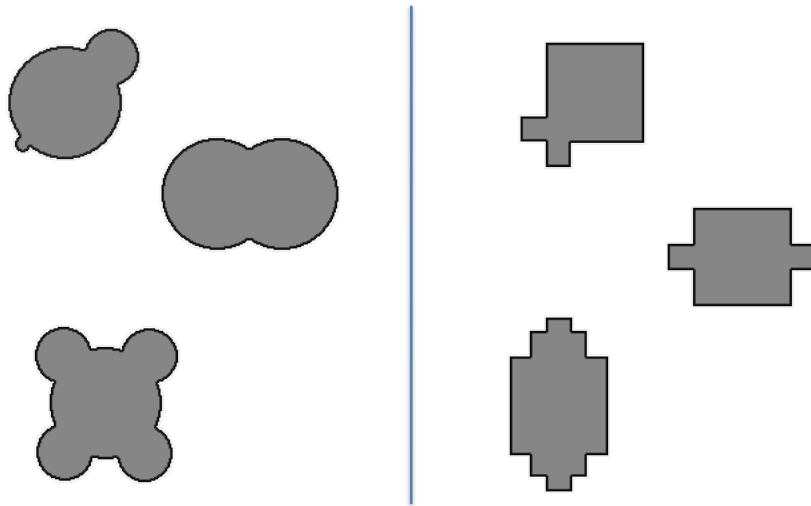


Figure 6.1: All 6 objects used in the experiment. The objects fall into 2 groups on the basis of whether they have curved or straight outlines. In each run of the experiment, one of these groups is designated as constrained and the other as unconstrained. The assignment of groups is counterbalanced across participants (see section 6.3.2.2).

expectations about the co-occurrence patterns of particular objects and actions. In the second phase, participants were trained on a 2-word language for 2 events drawn from the possible set. In the third phase, participants played a communication game in pairs, where their task alternated between labelling an event for their partner, and picking an event from an array based on a label provided by their partner. In the fourth phase, participants were individually tested on the language they had developed with their partner during the communication phase. In the final phase, participants were shown a series of events and asked to judge whether they were possible or impossible in the artificial world.

6.3.1 Participants

Participants were 48 students at the University of Edinburgh (37 female, median age 21). The experiment took 1 hour; participants were paid £7. In addition, each member of the pair with the highest score, calculated by dividing their communicative success by the time taken to complete the communication phase, was awarded a £10 Amazon voucher.

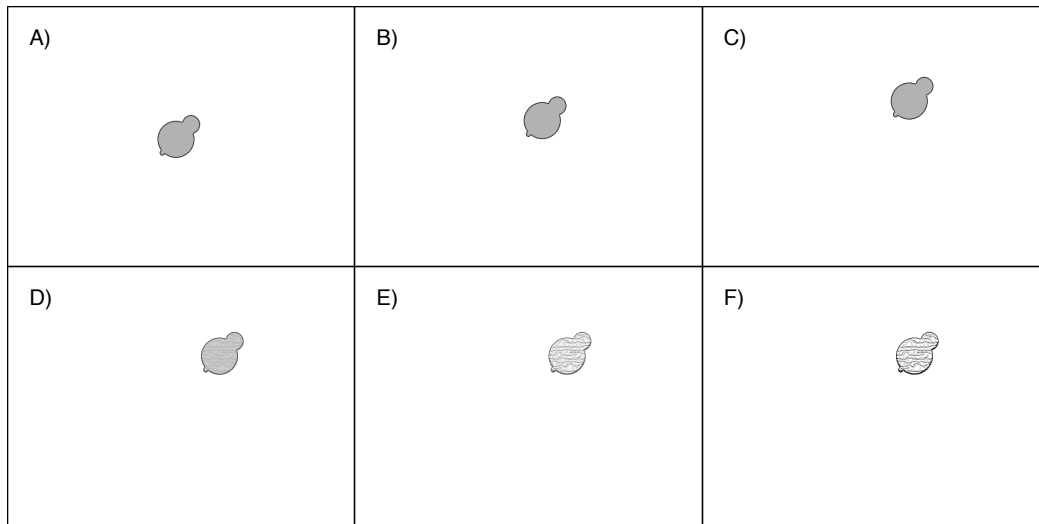


Figure 6.2: Static frames illustrating an animation stimulus from the experiment. First the object moves diagonally up and right (move action 1: A-C). Then object fill colour changes from grey to textured (fill action 2: D-F). Each action takes 3 seconds, giving an overall event duration of 6s. Examples of the animations are provided with the electronic version of the thesis.

6.3.2 Stimuli

The stimuli were animations depicting events, based on those used in Pirog Revill et al. (2008). In each event, objects performed two actions in sequence. Each event featured one of 6 possible objects (Figure 6.1) performing one of 3 possible ‘move’ actions and one of 3 possible ‘fill’ actions. ‘Move’ actions were: 1) object moves diagonally up and to the right; 2) object moves horizontally to the right and then to the left; 3) object moves in a spiral. ‘Fill’ actions were: 1) object changes colour from grey to white; 2) object changes colour from grey to textured; 3) object changes colour from grey to black. Figure 6.2 shows still images from one of the stimuli. ‘Fill’ and ‘move’ actions took 3s each, giving each stimulus a total duration of 6s. Examples of the stimuli are provided with the electronic version of the thesis.

6.3.2.1 Event structure constraints

The co-occurrence of objects and actions in events was non-random: certain combinations of objects and actions were possible in the artificial world, while others were not. Specifically, one class of objects and one class of actions were **constrained**, while the other class of objects and class of actions were **unconstrained**.

Object	Action 1	Action 2	Object	Action 1	Action 2
CO1	CA1	UA1	UO1	CA1	UA1
CO1	CA1	UA2	UO1	CA1	UA2
CO1	CA1	UA3	UO1	CA1	UA3
CO1	CA2	UA1	UO1	CA2	UA1
CO1	CA2	UA2	UO1	CA2	UA2
CO1	CA2	UA3	UO1	CA2	UA3
CO1	CA3	UA1	UO1	CA3	UA1
CO1	CA3	UA2	UO1	CA3	UA2
CO1	CA3	UA3	UO1	CA3	UA3
CO2	CA1	UA1	UO2	CA1	UA1
CO2	CA1	UA2	UO2	CA1	UA2
CO2	CA1	UA3	UO2	CA1	UA3
CO2	CA2	UA1	UO2	CA2	UA1
CO2	CA2	UA2	UO2	CA2	UA2
CO2	CA2	UA3	UO2	CA2	UA3
CO2	CA3	UA1	UO2	CA3	UA1
CO2	CA3	UA2	UO2	CA3	UA2
CO2	CA3	UA3	UO2	CA3	UA3
CO3	CA1	UA1	UO3	CA1	UA1
CO3	CA1	UA2	UO3	CA1	UA2
CO3	CA1	UA3	UO3	CA1	UA3
CO3	CA2	UA1	UO3	CA2	UA1
CO3	CA2	UA2	UO3	CA2	UA2
CO3	CA2	UA3	UO3	CA2	UA3
CO3	CA3	UA1	UO3	CA3	UA1
CO3	CA3	UA2	UO3	CA3	UA2
CO3	CA3	UA3	UO3	CA3	UA3

Table 6.1: Table showing possible and impossible events in the artificial world of the experiment. A row corresponds to an event. CO, constrained object; UO, unconstrained object; CA, constrained action; UA, unconstrained action. White rows are possible events, shaded rows are impossible events.

The combinations of objects and actions that made up the full set of 12 possible events are highlighted in Table 6.1. The constraints were as follows:

Constrained objects. Each constrained object appears with only 1 out of the 3 possible types of each action. For example, if the object shown in Figure 6.2 is a constrained object, it will only ever move diagonally up and right (never left & right or spiral), and only ever fill with texture (never white or black). The two remaining curved objects would be constrained to the two remaining move and fill actions, respectively. A real-world parallel to constrained objects is non-animate objects, which can take part in a smaller range of actions than animates can.

Constrained actions. Each constrained action appears with only one object of each type. For example, if move actions are constrained, then one object of each type is matched consistently with one move action. For example, the top curved object and the top straight object in Figure 6.1 would always move diagonally; the middle curved object and the middle straight object would always oscillate right and left; and the bottom curved object and the bottom straight object would always move in a spiral. A real-world parallel to constrained actions is selectional restrictions, or the requirement for particular actions to have particular participants: for example, the action of drinking requires a liquid patient.

Unconstrained objects and unconstrained actions. Each unconstrained object can appear with any unconstrained action. Thus, if fill is unconstrained, each unconstrained object could potentially turn white, textured or black.

It should be clear from these examples that these constraints are interdependent. Unconstrained objects are still constrained when it comes to constrained actions, and unconstrained actions are still constrained when it comes to constrained objects. However, across the whole set, this leads to differences in relative predictive value for constrained versus unconstrained objects and actions. Table 6.2 summarises these differences.

From the table we can see that the predictive value of each event feature depends on which other features are also known. For example, unconstrained action is a helpful predictive cue only if we know a) whether the object involved is constrained or unconstrained, b) if unconstrained, which specific object it is. The implications of these event structure constraints for the potential strategies participants might use in

Known	Predicts		
	Object	CA	UA
CO	-	✓	✓
UO	-	✓	X
CA	Narrows down to 2: one CO, one UO	-	✓ (if CO) X (if UO)
UA	✓ (if CO) X (if UO)	✓ (if CO) X (if UO)	-

Table 6.2: Table showing predictive value of event features. CO = constrained object, UO = unconstrained object, CA = constrained action, UA = unconstrained action.

the experiment are discussed in section 6.4.

6.3.2.2 Counterbalancing

The assignment of object groups (curved/straight) and action types (move/fill) as constrained or unconstrained was counterbalanced across participants. The order of the actions (constrained followed by unconstrained, or vice versa) was also counterbalanced across participants. For each pair, action order was consistent throughout the experiment, e.g. objects would always move first and then fill. The counterbalancing was intended to control for any effects of higher salience of one action type over another, or of the first action over the second action. Individual features, i.e., which particular objects were constrained to which particular actions, were assigned randomly for each pair.

6.3.3 Training language

The training language consisted of two words that labelled two events. These two events were randomly selected from the full set of possible events, within the constraints that a) the two events shared no features and b) one event involved a constrained object and one an unconstrained object. For example, the first and final rows in Table 6.1 form a potential training set.

The two words of the training language were randomly generated for each pair. Each word was 5 syllables long. This length was chosen to prevent participants in-

ferring a clear correspondence of syllable(s) to meaning feature(s), since every event had 3 features (the object, the type of movement, and the colour of fill). The labels were generated by concatenating randomly selected syllables from a set of 10 ('jo', 'xu', 'qi', 'ta', 'ru', 'wu', 'ye', 'su', 'mo', 'gi'). These syllables in turn were generated by randomly selecting consonants and vowels from the alphabet. The two labels each used mutually exclusive combinations of 5 syllables from the 10-syllable set, ensuring that the two labels did not share any elements at the syllable level or higher.

6.3.4 Procedure

Participants were told they would be learning about events in an alien world. Pairs of participants completed the experiment on computer terminals in separate cubicles. The familiarisation, training, post-test and event test phases were done individually, while the communication phase was done as a pair, communicating via the computers.

6.3.4.1 Familiarisation phase

Participants were shown the full set of possible events (white cells in Table 6.1). They were instructed to watch carefully and try to get an idea of which events were possible and impossible in the alien world. They were told their partner would be watching the same events.

To balance the overall occurrences of each object and each action, events involving constrained objects were shown three times each, while events involving unconstrained objects were shown once each. This meant that overall, each object appeared 3 times and each action appeared 6 times. Each event was preceded by a fixation cross that appeared for 1s, to make it clear where each event ended and the next began.

6.3.4.2 Training phase

Participants were told that they would learn some ways the aliens had of talking about the events they had just seen. They were told that their partner would be learning the same ways of talking about the same events. 'Ways of talking' was used rather than 'words' in order to leave open the possibility of segmenting the labels. On each trial, the participant was shown the event along with its label. The label then disappeared. They were shown the event once more and then asked to retype the label they had seen, pressing Enter when they were finished. The event continued to loop until they had pressed Enter, with a fixation cross showing for 1s at the end of each loop. There

were four rounds of training, with each round consisting of two trials: one for each of the two event/label combinations in shuffled order.

6.3.4.3 Communication phase

The participant who finished the training phase first was presented with a ‘Waiting for partner’ screen until their partner had completed training. The communication phase then began. The participants were told that their task was to use the alien language to communicate events to each other. Participants alternated roles between sender and receiver. On each trial, the sender was shown a single event from the set of possible events. The sender was asked to type a message in the alien language to convey this event to their partner. The sender could type any letters in the alphabet and could also use spaces. Once the sender had typed a label and pressed Enter, the receiver was shown the label the sender had typed and an array of four events: the target event, and three distractors. These distractors were drawn at random from the set of possible events in the alien world; participants were informed in the instructions that this would be the case. Once the receiver had picked an event, both participants were shown a feedback screen. This showed the word the sender typed, the target event, the event the receiver picked, whether communication was successful or unsuccessful, the score for the trial (1 if communication was successful, 0 if it was unsuccessful), and the total score for the communication phase so far. Participants then swapped roles for the next trial, i.e., the receiver became the sender and the sender became the receiver.

Communication consisted of two rounds. Each round contained 18 trials. As in the familiarisation phase, these 18 trials consisted of 3 instances of each event involving a constrained object, and 1 instance of each event involving an unconstrained object. This was done to balance the overall frequency of occurrence of each object and each action, such that each object appeared 3 times and each action appeared 6 times overall. The order of trials was randomised, within the constraint that each participant acted as the sender once and the receiver once for each of the events in the set over the 2 rounds. For each event involving a constrained object, each participant acted as the sender three times and the receiver three times.

6.3.4.4 Post-test phase

Participants were shown whole events and parts of events and asked to type what they would call them in the language they used with their partner in the communication

phase. They were instructed that if they did not know what to call an event, they could press Enter without typing anything.

The stimuli shown in the post-test were as follows:

- All possible events (i.e., the same set shown in the familiarisation and communication phases, but with each event shown only once)
- Events involving a novel object (a square with rounded corners) performing every combination of move and fill actions
- Single actions (i.e., moving alone or filling alone) performed by objects for which those actions were possible during the experiment
- Single actions (i.e., moving alone or filling alone) performed by a novel object
- Objects alone, not performing an action

The set of possible events from the experiment was presented in order to extract a ‘clean’ version of the language the participants developed during the communication phase. The rest of the items in the post-test set were included to allow for unambiguous detection of whether participants had labels for each action (if they labelled the single actions performed by a novel object) and for each object (if they labelled the objects in isolation).

Stimuli were presented in shuffled sets of decreasing event complexity: first, all fully complex events involving two actions; second, all events involving a single action, where that action type was the one that was habitually shown first during the experiment; third, all events involving the second action type; and last, all objects in isolation. This made 51 trials in total. Unlike in the familiarisation and communication phases, the frequency of individual objects and individual actions was not balanced in the post-test. This was necessary in order to show all possible events and sub-events, while also not making participants label the same event more than once.

6.3.4.5 Event test phase

In the final phase of the experiment, participants were shown events and asked to judge, based on what they had seen during the course of the experiment, whether each event was possible or impossible in the alien world. On each trial, the participant was shown an event and asked to press Y if the event was possible and N if the event was impossible. The participant could interrupt the event at any time to press Y or N. The event

Object	Action 1	Action 2
CO1	CA2	UA1
CO2	CA1	UA2
CO3	CA3	UA1
CO1	CA1	UA3
CO2	CA3	UA3
CO3	CA2	UA2
UO1	CA2	UA1
UO1	CA3	UA2
UO2	CA1	UA3
UO2	CA3	UA1
UO3	CA1	UA2
UO3	CA2	UA3

Table 6.3: Table showing combinations of objects and actions making up the test set of impossible events in the event test phase of the experiment. Features that violate event structure constraints are shown in bold. A row corresponds to an event. CO, constrained object; UO, unconstrained object; CA, constrained action; UA, unconstrained action.

looped continuously until Y or N was pressed, with a fixation cross displayed for 1s at the end of each loop. Once Y or N was pressed, the next trial started. Participants were not given feedback as to whether their possible/impossible judgement was correct or incorrect.

The event test phase had 24 trials in total. 12 trials featured each of the 12 possible events from Table 6.1, as used in the familiarisation and communication phases, but with each event presented only once. The other 12 trials featured events that violated the event structure constraints of the artificial world in various ways. These events are shown in Table 6.3, with the features that violate event structure constraints shown in bold. The order of trials was shuffled. While the frequency of particular objects and actions was not balanced within the set of possible events, due to the nature of the co-occurrence constraints, these frequencies were balanced within the set of impossible events. Half of the impossible events involved constrained objects, and half involved unconstrained objects. Of the events involving constrained objects, 2 violated event structure constraints by pairing them with the wrong constrained action; 2 by pairing

them with the wrong unconstrained action; and 2 by pairing them with both the wrong constrained action and the wrong unconstrained action. For the events involving unconstrained objects, all involved pairing them with the wrong constrained action (since this was the only constraint for these objects).

6.3.5 Dependent variables

6.3.5.1 Event test score

Participants' scores on the event test were recorded to determine whether they had successfully learned the event structure constraints in the artificial world. The score for the event test was the number of trials for which the participant correctly judged that an event was possible or impossible. Where a participant was incorrect, whether this was a false positive (i.e., judging an impossible event to be possible) or a false negative (i.e., judging a possible event to be impossible) was also recorded. d' , a signal detection measure based on the proportion of hits (Y answers to possible events) to false alarms (Y answers to impossible events), was also calculated.

6.3.5.2 Communicative success

Participants' communicative success scores were measured to determine whether they were above chance levels and whether they improved across rounds. Communicative success was the sum over the communication phase of the 0 or 1 scores awarded for success or failure on each trial.

6.3.5.3 Event feature encoding

The main variable of interest is the extent to which event features (objects and actions) are encoded in the labelling patterns produced during communication and in the post-test. We therefore need a way of quantifying the extent to which each feature is expressed in the language. However, since participants can freely type their labels and are not instructed to use spaces or mark which part(s) of a label correspond to which part(s) of an event, we have the double challenge of segmenting the labels produced, and determining which features are strongly and consistently expressed by these segments. The constrained co-occurrence structure of the stimuli further complicates the task of finding unique associations for each feature.

To deal with these problems, a novel analysis method was applied. The data to be analysed was each participant's labelling data, isolated by taking from the communication phase only those trials where they were the sender. The post-test phase was already separated by participant. An analysis was run on all the label-stimulus pairs from each phase to find a) which sub-strings were associated with which meaning features, and b) how strong or consistent the association was. The associations accepted between meaning features m and signal elements s are those that satisfy the following two constraints: 1) When m appears in the target stimulus, s shows up in the label more consistently than any other signal element; 2) s is more reliably associated with m than with any other meaning feature. To determine and quantify these associations, the following analysis was performed:

1. Stimuli were coded in terms of meaning features, as in Table 6.1. Labels were split into all possible constituent n-grams: e.g., the label 'feju' contains the n-grams 'f', 'e', 'j', 'u', 'fe', 'fej', 'ej', 'eju', 'ju', and 'feju'.¹
2. A matrix where columns corresponded to meaning features and rows corresponded to n-grams was populated with association values for each meaning feature and each n-gram. These values were the number of co-occurrences of a meaning feature m and an n-gram s , divided by the total number of occurrences of m . These values ranged from 0 to 1, with 0 indicating that s is never in the label for stimuli that include m , and 1 indicating that s is always in the label for stimuli that include m . For each meaning feature m , the n-gram(s) were ranked from highest to lowest association strength with m .
3. Each n-gram with the maximum association strength with m is then checked to see if it is more associated with m than with any other meaning feature. If any other meaning feature has higher or equal² association with this n-gram,

¹Only adjacent n-grams were included in the analysis: for example, for 'feju', 'f*ju' etc. were not included as candidate n-grams. An analysis that also included all possible non-adjacent n-grams was found to be computationally intractable, given that participants used labels of up to 21 letters in length. However, where consistent labels for features can be identified from a descriptive analysis, they tend to be contiguous rather than split, supporting the exclusion of non-adjacent n-grams from the analysis.

²This competitive analysis takes into account the constrained co-occurrence structure of the objects and actions. For example, in the communication phase, each object o always appears with the same constrained action c , whereas this constrained action appears with this object and one other. Thus, stimuli containing o are a subset of stimuli containing c . If an n-gram s therefore has an equal association with o and c , this association is accepted for c and rejected for o . For example, in the language in Figure 6.3, 'ju' has equal 1.0 associations with CO1, UO1 and CA1; however, because stimuli containing CO1 and UO1 are a subset of the stimuli containing CA1, this association is accepted for CA1 while being rejected for CO1 and UO1.

the association is not accepted for m . The next-highest associated n-gram(s) are then tested in the same way, until an n-gram or n-grams are found which have the highest level of association unique to m .

4. If multiple candidate n-grams fit these criteria, then the n-gram that includes the most other candidate n-grams is accepted as the label: e.g., if ‘f’, ‘fe’, ‘ej’, ‘fej’, ‘eju’, and ‘feju’ all have the same level of unique association with m , ‘feju’ is accepted as the label.
5. Each meaning feature now has a label and a corresponding association strength. The final step loops through each of these labels: if a label for a feature m appears within a label for another feature n and its association with m is higher than or equal to its association with n , the label for m is removed from the label for n . If this leaves the label for n empty, the association strength for n is set to 0.

Figure 6.3 shows an example of this analysis performed for the meaning feature UO1, on a dummy language that encodes every feature. The analysis returns a label for each meaning feature, along with an association strength. This latter value, reported in all analyses below, serves as a quantitative measure of the extent to which each meaning feature is conventionalised in the language.

The associations that come out of this analysis tally well with a descriptive analysis of the languages. An additional advantage is that since the analysis normalises for the number of occurrences of each meaning feature, we can compare objects with actions (despite each action having more occurrences than each object) and constrained objects and actions with their unconstrained equivalents (despite different frequencies of occurrence during the post-test).

To test the validity of this analysis, it was run on constructed languages encoding different combinations of features. For each language tested, the analysis correctly extracted the part of the label that encoded each feature and gave an association strength of 1. For non-encoded features, the analysis returned association values of 0.

6.4 Hypotheses

6.4.1 Predictable versus unpredictable

The main hypothesis is that unpredictable event features will tend to be lexicalised more strongly than predictable event features, i.e., the association strength values from

1. Labelling data:

CO1CA1UA1	"bajuma"	UO2CA2UA1	"gakoma"
CO2CA2UA2	"cikono"	UO2CA2UA2	"gakono"
CO3CA3UA3	"dulepi"	UO2CA2UA3	"gakopi"
UO1CA1UA1	"fejuma"	UO3CA3UA1	"dulema"
UO1CA1UA2	"fejuno"	UO3CA3UA2	"duleno"
UO1CA1UA3	"fejupi"	UO3CA3UA3	"dulepi"

2. N-grams with highest association with UO1:

"f", 1.0	"ej", 1.0
"e", 1.0	"ju", 1.0
"j", 1.0	"fej", 1.0
"u", 1.0	"eju", 1.0
"fe", 1.0	"feju", 1.0

3. Competitive analysis:

"f" is more associated with UO1 than any other feature. **Candidate**
 "e" is equally associated with CO3 and UO3. Rejected
 "j" is equally associated with CO1 and CA1. Rejected
 "u" is equally associated with CO1, CO3 and CA1. Rejected
 "fe" is more associated with UO1 than any other feature. **Candidate**
 "ej" is more associated with UO1 than any other feature. **Candidate**
 "ju" is equally associated with CO1 and CA1. Rejected
 "fej" is more associated with UO1 than any other feature. **Candidate**
 "eju" is more associated with UO1 than any other feature. **Candidate**
 "feju" is more associated with UO1 than any other feature. **Candidate**

Candidates at this stage: "f", "fe", "ej", "fej", "eju", "feju"

4. Maximise n-gram inclusions:

"feju" includes all other candidate n-grams. "feju" wins

5. Remove n-grams with winning associations with other features:

"ju" wins for CA1 with association 1.0. Remove from string for UO1

FINAL RESULT: "fe", 1.0

Figure 6.3: Extracting a label and corresponding association strength for the meaning feature UO1 (unconstrained object 1) from communication data. In this constructed language, 'fe' is used consistently for UO1. The figure shows step-by-step how the analysis extracts the association 'fe' for UO1 from the data. Because this constructed language is completely consistent, UO1 has an association strength of 1.0, the maximum value.

the feature encoding analysis will be higher for unpredictable event features.

Table 6.4 gives two examples of constructed languages which lexicalise each object and either constrained or unconstrained actions. As the table shows, lexicalising unconstrained actions gives a more expressive language (i.e., one that ensures efficient inference of meaning from linguistic cues and world knowledge) for the same number of labels. Therefore, we can predict that unconstrained actions should be lexicalised more strongly than constrained actions, since they are a more efficient predictive cue overall. This prediction is linked to the broad hypothesis that learning and communication pressures combine to create a compromise between expressivity and learnability: the languages will lexicalise only those features that do not come for free, i.e., cannot be inferred on the basis of other lexicalised features.

For constrained versus unconstrained objects, the picture is less clear. As stated in the caption to Table 6.4, in a language that lexicalises unconstrained actions, unconstrained objects must still be lexicalised to gain full expressivity. However, for constrained objects, it is possible to predict all features of the event by using the unconstrained action label either alone or with an object class marker. We might therefore also expect constrained objects to be lexicalised less strongly than unconstrained objects. However, as pointed out in the Introduction, what is predictable in an utterance context depends not only on event structure constraints, but also on which event features are already lexicalised in a language. Thus, we can expect emerging patterns of lexicalisation to be constrained by the conventions that are established early in communication. If constrained objects are lexicalised early, it may be less costly to keep these conventions, even if they result in redundancy. Overall, lexicalisation patterns are expected to be a result of the interplay between early conventions, communicative efficiency, and the generalisation of patterns across the language.

6.4.2 Objects versus actions

We also have a priori reasons for expecting objects to be lexicalised more strongly than actions, regardless of their constrained or unconstrained status. Referring back to section 3.5.1 of Chapter 3, this prediction comes from the work of Dedre Gentner, who argues that relational categories (e.g., spatial relations and actions) are cognitively less accessible than object categories (Gentner & Kurtz, 2005). As elaborated by Croft & Cruse (2004), actions may be more difficult to abstract from specific instances, since they are dependent on their arguments: a movement cannot be conceptualised without

Object	Action 1	Action 2	Label
CO1	<i>CA1</i>	<i>UA1</i>	baju
CO2	<i>CA2</i>	<i>UA2</i>	ciko
CO3	<i>CA3</i>	<i>UA3</i>	dule
UO1	<i>CA1</i>	<i>UA1</i>	feju
UO1	<i>CA1</i>	<i>UA2</i>	feju
UO1	<i>CA1</i>	<i>UA3</i>	feju
UO2	<i>CA2</i>	<i>UA1</i>	gako
UO2	<i>CA2</i>	<i>UA2</i>	gako
UO2	<i>CA2</i>	<i>UA3</i>	gako
UO3	<i>CA3</i>	<i>UA1</i>	dule
UO3	<i>CA3</i>	<i>UA2</i>	dule
UO3	<i>CA3</i>	<i>UA3</i>	dule
Object	Action 1	Action 2	Label
CO1	<i>CA1</i>	<i>UA1</i>	bama
CO2	<i>CA2</i>	<i>UA2</i>	cino
CO3	<i>CA3</i>	<i>UA3</i>	dupi
UO1	<i>CA1</i>	<i>UA1</i>	fema
UO1	<i>CA1</i>	<i>UA2</i>	feno
UO1	<i>CA1</i>	<i>UA3</i>	fepi
UO2	<i>CA2</i>	<i>UA1</i>	gama
UO2	<i>CA2</i>	<i>UA2</i>	gano
UO2	<i>CA2</i>	<i>UA3</i>	gapi
UO3	<i>CA3</i>	<i>UA1</i>	duma
UO3	<i>CA3</i>	<i>UA2</i>	duno
UO3	<i>CA3</i>	<i>UA3</i>	dupi

Table 6.4: Two languages that lexicalise each object (bold) and one action class (italic). The upper language lexicalises each object and the constrained action. The lower language lexicalises each object and the unconstrained action. The upper language would not be optimal for communication: it independently lexicalises features that are predictable from each other (unconstrained object and constrained action), but does not lexicalise an unpredictable feature (unconstrained action). The lower language would be optimal for communication. Out of the two action classes, the unconstrained action is more adaptive to lexicalise. Both languages contain some redundancy: for events involving constrained objects, either identifying the object alone (i.e. 'ba', 'ci', 'du'), using a class label plus the unconstrained action (i.e., 'bama', 'bano', 'bapi'), or an 'unmarked' use of the unconstrained action label alone ('ma', 'no', 'pi') would be sufficient to infer the whole event.

something that is moving, whereas an object can readily be conceptualised in isolation. We may therefore expect higher lexicalisation rates for objects than actions in the experiment. Potentially balancing this, however, is the fact that in the experiment as a whole and in each phase, each object shows up fewer times than each action; therefore, while individual object labels may be helpful predictive cues, each one can be applied on fewer trials than each individual action label. These factors may interact over the course of the experiment to lead to eventual equivalent lexicalisation rates for objects and actions.

6.5 Results

6.5.1 Event structure test

The average score in the event structure test was 16 out of 24 ($SD = 3.72$). A one-sample Wilcoxon signed-rank test found this was significantly above chance, $z = -5.15$, $p < .001$. Figure 6.4 shows a d' analysis, calculated by subtracting the z-score transformation of participants' false positives (proportion of impossible trials where they responded Y) from participants' hits (proportion of possible trials where they responded Y). Participants' average d' score was 1.27, 95% CI [0.93, 1.61], showing that on average participants performed significantly above the chance value of 0, $z = -5.14$, $p < .001$.

However, while on average participants were above chance, the distribution of scores was bimodal: around a third of participants scored around chance (10-14 points), with the remaining participants achieving scores between 15 and 24. This suggests that the first group of participants either did not acquire the event structures they were trained on, or did not understand the task. Specifically, 5 participants scored 0 on the d' analysis because they pressed 'Y' for every event, regardless of whether it was possible or impossible. More generally, false positives (pressing 'Y' for an event that was not possible) were more common than false negatives (pressing 'N' for an event that was possible), with the average participant making 76% false positive errors and 24% false negative errors ($SD = 20\%$). The implications of this are explored in section 6.6.1 below.

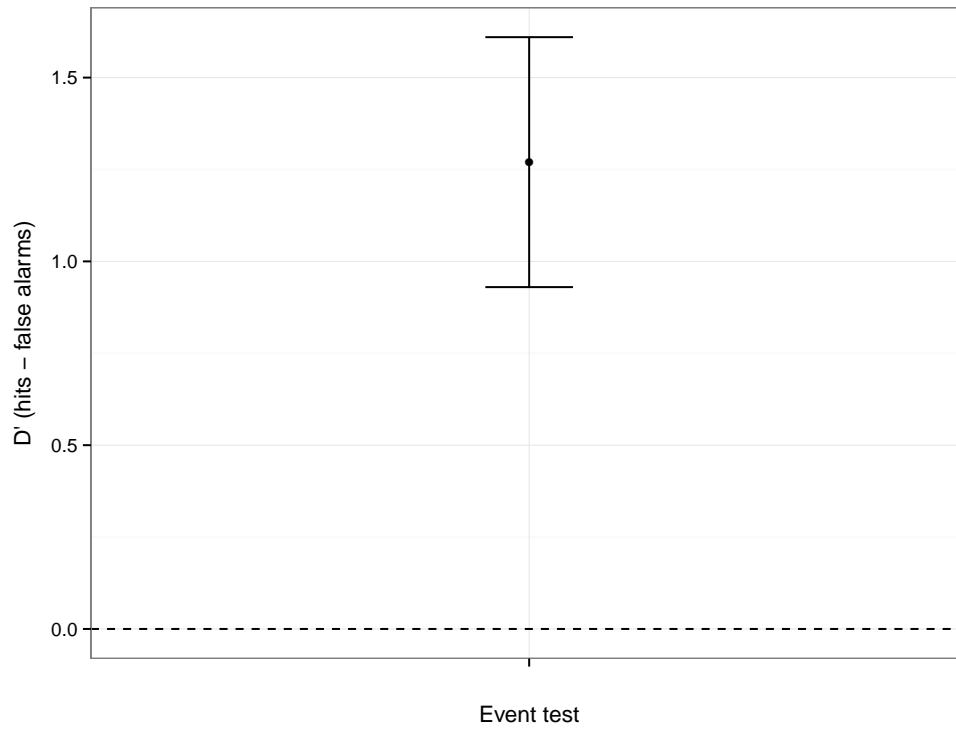


Figure 6.4: Participants' responses in the event test coded as d' scores. d' is the difference between the z -score transformations of each participant's hits (i.e., proportion of possible event trials where they responded Y) and false alarms (i.e., proportion of impossible event trials where they responded Y). The dotted line at 0 shows chance: participants who score 0 have the same number of hits as false positives. This includes 5 participants who responded Y to every event (see text). Error bar is the 95% confidence interval.

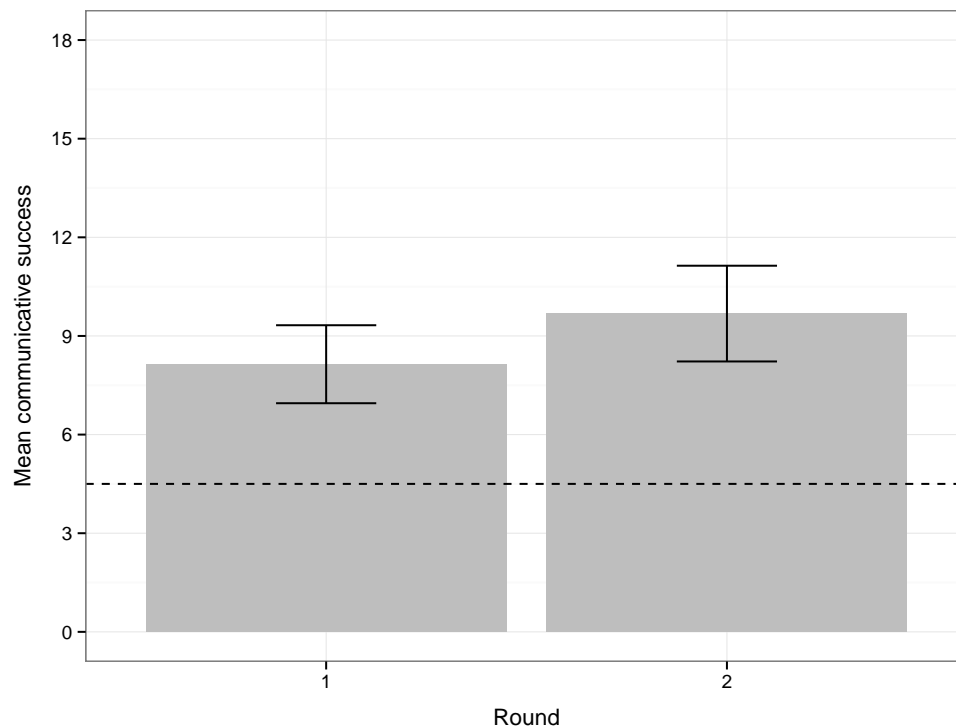


Figure 6.5: Communicative success scores over rounds in the experiment. Dotted line shows chance per round (4.5 out of 18). Error bars are 95% confidence intervals.

6.5.2 Communicative success

Two pairs used some English in their language (e.g., including the letter ‘b’ in labels for stimuli where the fill action was black). These participants were excluded from the analysis of communicative success, since this may have artificially boosted their score. Average success for the remaining participants was 18 [15.69, 19.95], corresponding to 50% of the maximum score of 36. However, since chance success was only 25%, participants’ performance was significantly above chance, $t(21) = 8.62, p < .001$. Figure 6.5 shows the change in success scores across rounds, with the dotted line showing chance score per round.

Average score in round 1 was 8/18 [6.95, 9.33]. Average score in round 2 rose to 9/18 [8.22, 11.14], an increase of 1.55 [-0.05, 3.14]. As the confidence interval suggests, the difference between the scores in the two rounds was marginally significant, $t(21) = 2.02, p = .057$. Therefore, while pairs did improve from round 1 to round 2, this increase was not numerically large – around 1 extra successful trial out of 18, on average. However, Cohen’s unbiased d for the increase was 0.63, suggesting a medium effect.

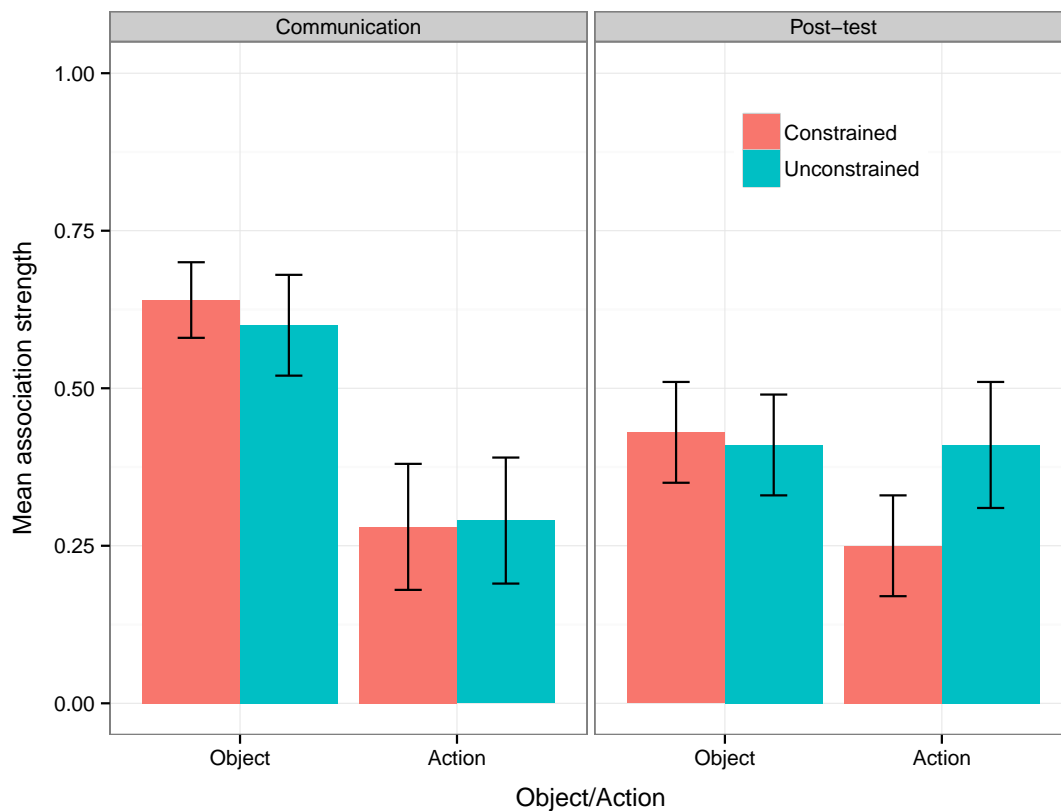


Figure 6.6: Mean association strength for each class of object and action during the communication and post-test phases of the experiment. Error bars are 95% confidence intervals.

6.5.3 Event feature encoding

The analysis described in section 6.3.5.3 was run to determine which features were more or less strongly lexicalised in the output languages. The results were analysed using mixed-effects models. For all the models in this section, as in those for Experiment 1 in Chapter 4, random effects for inclusion were assessed by likelihood ratios, i.e. testing models against each other to see if additional parameters improved model fit. p -values for the analyses were estimated using Baayen's formula, with the additional heuristic of only accepting absolute t -values greater than 2 as significant.

Figure 6.6 shows the results for association strength over the whole experiment. During the communication phase, constrained objects had a mean association strength of 0.64 [0.58, 0.70].³ Unconstrained objects had a mean association strength of 0.60

³All confidence intervals are calculated based on association values averaged across the three representatives of each category. For the communication phase, association values are also averaged across members of each pair, since these observations were not independent.

[0.52, 0.68]. This was a difference of 0.04 [-0.03, 0.09]. Constrained actions had a mean association strength of 0.28 [0.18, 0.38], while unconstrained actions had a mean association strength of 0.29 [0.19, 0.39]. This was a difference of 0.01 [-0.16, 0.18].

During the post-test phase, constrained objects had a mean association strength of 0.43 [0.35, 0.51]. Unconstrained objects had a mean association strength of 0.41 [0.33, 0.49]. This was an average difference of 0.02 [-0.04, 0.08]. Constrained actions had a mean association strength of 0.25 [0.17, 0.33], while unconstrained actions had a mean association strength of 0.41 [0.31, 0.51]. The difference was 0.16 [0.04, 0.28].

These differences were analysed using a linear mixed-effects model. The model incorporated a random slope for Participant by a three-way interaction of Object/Action, Constrained/Unconstrained, and Communication/Post-test: i.e., the model took into account the variance between participants in how their association strength scores changed in response to all these effects and their interactions. Likelihood ratio tests indicated that including these random effects produced the best-fitting model. The model included fixed effects of Object/Action, Constrained/Unconstrained, and Communication/Post-Test, and a three-way interaction.

The model found a significant main effect of Object/Action, $\beta = 0.36$, $t = 7.28$, $p < .001$. The main effect of Constrained/Unconstrained was not significant, $\beta = 0.01$, $t = 0.21$, $p = .83$. Neither was the main effect of Communication/Post-test, $\beta = -0.02$, $t = -0.53$, $p = .59$. The interaction between Object/Action and Constrained/Unconstrained was not significant, $\beta = -0.04$, $t = -0.57$, $p = .57$. However, the interaction between Object/Action and Communication/Post-test was significant, $\beta = -0.20$, $t = -2.95$, $p = .003$. So was the interaction between Constrained/Unconstrained and Communication/Post-test, $\beta = 0.15$, $t = 2.47$, $p = .01$. The three-way interaction did not quite reach significance, $\beta = -0.14$, $t = -1.76$, $p = .08$.

To unpack these effects with reference to Figure 6.6: objects overall had higher association strengths than actions. Association strengths were similar on average during the communication and post-test phases. However, the interaction between Object/Action and Communication/Post-test suggests that the difference between objects and actions was less in the post-test than during communication. Overall, event structure constraints did not have a significant effect on association strengths; however, the interaction between Constrained/Unconstrained and Communication/Post-test suggests that the manipulation had an effect during one of these phases but not the other.

An examination of Figure 6.6 suggests that the interaction between Constrained/Un-

constrained and Communication/Post-test is driven by the lower association values for constrained actions versus unconstrained actions in the post-test. To investigate this further, post-hoc paired t -tests were run to compare constrained and unconstrained actions during communication and during the post-test. These post-hoc tests were performed on observations averaged across the three instances of each action class (also averaged across pairs for the communication phase), with alpha set to $p < .025$ according to the Bonferroni correction for multiple comparisons. The tests found no significant difference between constrained and unconstrained actions during the communication phase, $t(47) = -0.172, p = .86$, but a significant difference between constrained and unconstrained actions during the post-test phase, $t(47) = -2.53, p = .01$. d_{umb} for this latter difference was 0.5, suggesting a small to medium effect.

6.6 Discussion

6.6.1 Event test

As shown by their average above-chance performance in the event test, participants were able to acquire the event structure constraints of the artificial world. However, as noted in the Results, the distribution of scores was bimodal, suggesting that participants were split between those who performed well and those who performed at chance. The question is whether participants who performed at chance did so because of poor learning, or because of misunderstanding the nature of the task. In particular, 5 participants scored at chance (12 points) but had only false positive errors, i.e., they accepted every event they were shown as possible.

There are a number of potential explanations for this. These participants might have simply failed to learn the constraints; alternatively, they might have made the same response on every trial in order to finish the experiment faster; alternatively, they might have reasoned that just because they never saw a particular event during the experiment, that did not constitute sufficient evidence that it was impossible. The consistency of Y responses for these participants makes it unlikely that poor learning alone was responsible. These participants were also not notable outliers in the average time taken to press Y or N, speaking against explanation 2. More generally, there was no significant correlation between average time taken to guess and overall success in the event test, $r = .02 [-.26, .30], p = .89$. Overall, then, explanation 3 appears most likely. Rather than being poor learners, these participants instead appear not to have

taken the absence of particular co-occurrences as evidence that they were impossible.

A concern arising from this is that misunderstanding of the event constraints may have affected these participants' lexicalisation patterns. However, event test score did not correlate significantly with post-test differences in lexicalisation of constrained vs. unconstrained objects, $r = .02$ [-.27, .30], $p = .90$, or constrained vs. unconstrained actions, $r = 0.10$ [-.19, .38], $p = .49$. In addition, excluding these participants from the event feature encoding analysis did not change the statistical pattern of the results. As such, these participants may have formed probabilistic expectations about the likelihood of particular combinations, while not being willing to make binary possible/impossible judgements. Potential ways to test for this in future work are discussed in section 6.6.5.

6.6.2 Communicative success

The marginal improvement between rounds, and the fact that communicative success remained around 50%, showed that pairs did not necessarily develop fully expressive languages during communication. However, despite the difficulty of communicating after being trained on conventions for only 2 events, participants were able to succeed at levels above chance in the communication game.

6.6.3 Event feature encoding

The significant overall difference between object and action lexicalisation rates, along with the interaction with experimental phase, suggests that during communication, participants focused on conventionalising object labels. This supports the hypothesis that objects are, at least initially, more likely than actions to acquire stable associations.

However, during the post-test phase, the difference between objects and actions evened out. Constrained objects, unconstrained objects, and unconstrained actions all had roughly equivalent levels of association. However, constrained actions were lexicalised significantly less than unconstrained actions. Thus, the predicted difference between constrained and unconstrained actions did not emerge during communication, but only afterwards, when participants were asked to record and generalise the language they had used with their partner. This shows an interesting divergence between participants' communicative behaviour and their generalisation based on what they inferred during communication. Even though constrained and unconstrained actions appeared to be lexicalised at equivalent levels during the communication phase, par-

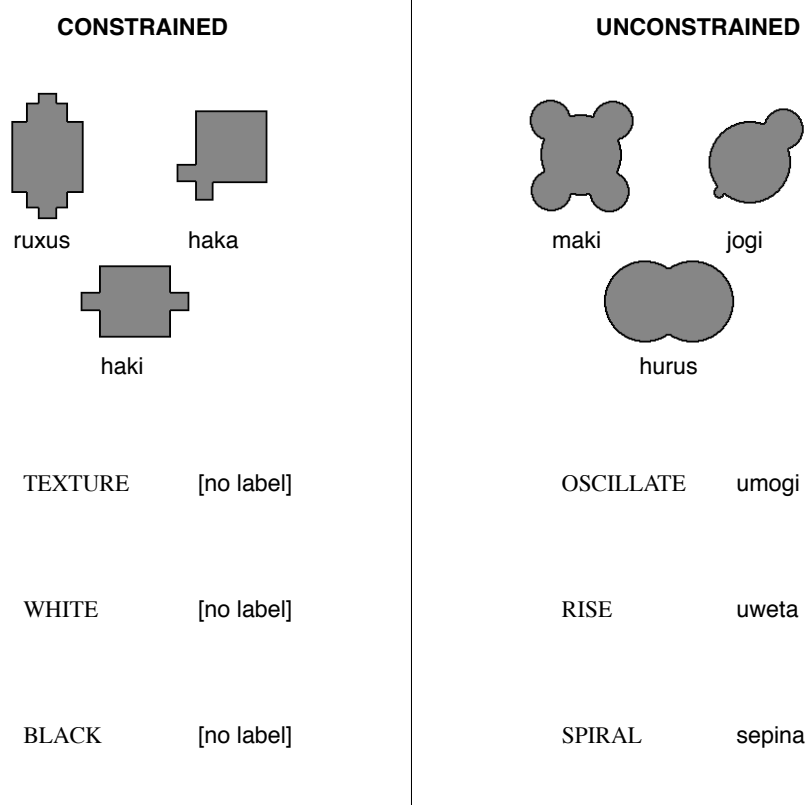


Figure 6.7: Example of a post-test language that lexicalised each object and each unconstrained action, but not the constrained actions. This language would allow for perfect success on the communication game, since the constrained action can always be inferred from one of the other lexicalised features. For example, ‘jogisepina’ specifies UO2 and UA3; from this, the receiver can infer that the other action must be CA2 (filling white), since UO2 never occurs with another constrained action. ‘ruxusumogi’ specifies CO1 and UA1; from this, the hearer can infer that CA1 (filling with texture) must be involved, since CO1 never occurs with a different constrained action. However, in this latter case, CO1 already predicts UA1, showing that the language contains some redundancy.

ticipants were more likely to infer patterns whereby unconstrained actions were more strongly lexicalised, and produce these patterns in their post-test data.

Turning to objects, the results suggest no significant difference in lexicalisation rates of constrained versus unconstrained objects, either during communication or in the post-test. This is potentially surprising since, as noted in Table 6.4, lexicalising constrained objects actually adds redundancy to a language that already lexicalises unconstrained actions. Figure 6.7 shows an example of a language from the post-test where unconstrained actions are lexicalised and constrained actions are not. If used consistently, this language allows perfect success in the communication phase, despite none of the constrained actions being lexicalised. However, it contains redundancy, since the unconstrained action already predicts which constrained object is involved in the event and vice versa. The most efficient language would use an unmarked unconstrained action label for events involving constrained objects, allowing for maximum expressivity with minimum number of labels to learn; however, from inspection of the output languages, no pair appeared to use this strategy. The next-most-efficient strategy would be to use a single object class label for constrained objects. The use of ‘haka’/‘haki’ as labels for two of the constrained objects in the language shown in 6.7 may show the beginning of a collapse in this direction. Section 6.6.4.1 below explores further the extent to which participants used object class label strategies.

6.6.4 Strategies

This section presents some examples of interesting phenomena in the results that are not brought out in the quantitative analysis.

6.6.4.1 Object class labels

One question raised in the previous section is whether participants might adopt a strategy of labelling object class (constrained vs. unconstrained, or, perceptually, curved vs. straight). Object class was included as a meaning feature in the analysis in order to determine whether this was a strategy participants used, and if so whether they would use it preferentially for constrained or unconstrained objects.

Object class labels were very rare: as such, a descriptive rather than a fully quantitative analysis follows. During the communication phase, 7 participants showed some evidence of using a label for the constrained object class. Of these, 4 had an association level of 1.0, 1 of 0.67, and 2 of 0.33. 8 participants showed some evidence of

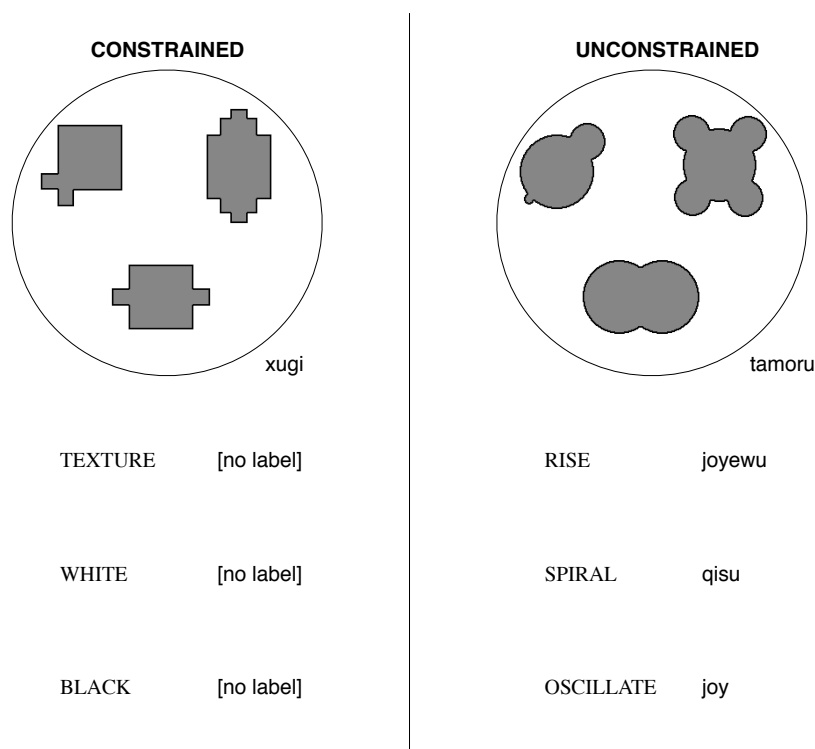


Figure 6.8: Example of a post-test language from a participant who used consistent object class labels. ‘joy’ is interpolated from a descriptive analysis: this label was not used consistently enough to be extracted by the quantitative analysis. This language would not guarantee perfect success on the communication game: for example, the label ‘tamorujoyewu’ is ambiguous for which curved object is involved, since all unconstrained objects can perform all unconstrained actions. For unconstrained objects, on the other hand, ‘xugijoyewu’ is not be ambiguous, since only one of the constrained objects can possibly perform each unconstrained action.

using a label for the unconstrained object class. Of these, 3 had an association level of 1.0, and 5 of 0.67. The number and level of association for object class labels in the communication phase therefore looks fairly similar for constrained and unconstrained objects.

During the post-test phase, 4 participants showed evidence of using a label for the constrained object class (2 with an association of 1.0, 1 of 0.5 and 1 of 0.25). 8 participants showed evidence of using a label for the unconstrained object class (4 at 1.0, with the remaining associations at 0.88, 0.75, 0.63, and 0.125). On balance, then, while frequency and association strength held steady for unconstrained object class labels during the post-test, it appeared to decline for constrained objects. However, the numbers are too small to be confident about this difference. Its direction is surprising, given the

point made in section 6.4: a system that used a class label for constrained objects and individual labels for unconstrained objects would be most efficient. However, given the high lexicalisation levels for individual constrained objects during communication, it is possible that once these conventions were established, dropping or modifying them to replace them with a class label was too costly for participants.

Figure 6.8 shows an example of a language that used object class labels for both constrained and unconstrained objects. As described in the legend, this language would be unambiguous for events involving constrained objects, but ambiguous for events involving unconstrained objects. An interesting possibility is that the participant has made a sub-optimal generalisation that the presence of one object class label implies that distinctions should be collapsed within the other object class, regardless of whether this is optimal for the language. Tentatively supporting this hypothesis is the fact that while the class label for the constrained objects already has an association value of 1.0 for this participant in the communication phase, the association for the unconstrained object class label is initially weaker (0.67 during communication), reaching 1.0 in the generalisation of the post-test. However, this is one case and so should be interpreted with caution.

6.6.4.2 Object labels derived from action labels

The experiment requires participants to generalise labels for whole events to individual object and action features. Often this involves some negotiation during the communication phase on whether a particular label or part refers to an object or an action. Evidence of these negotiations can be seen in the patterns of reuse of words in the post-test.

One participant produced in their post-test a language which only had labels for objects. However, these labels included recurring elements which derived from shared constrained action features. For example, the label for CO₂ was ‘rujotawas^u’, while the label for UO₂ was ‘rujotawux^u’; the label for CO₃ was ‘temprop^u’, while the label for UO₃ was ‘temp^u’. These pairs of objects have no perceptual features in common and are not in the same object class. However, they habitually perform the same constrained action. Therefore, while ‘rujota’ and ‘temp*^u’ are not generalised as action labels by this participant, their language shows evidence of this derivation in its object labels. An example of this from English is nouns such as director and rector, where the shared element refers to a shared habitual activity (ruling, presiding) that is not independently lexicalised by this label element in English.

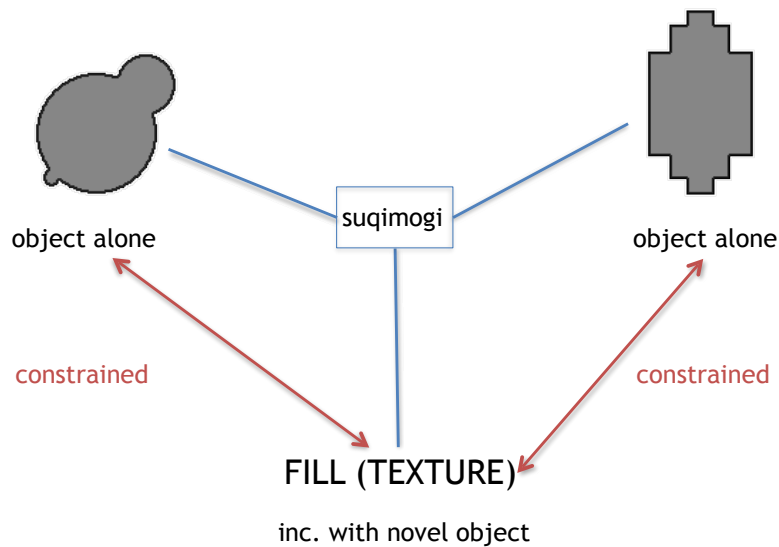


Figure 6.9: Example of a label generalised by a participant in the post-test phase to refer to a) both of the two objects that are constrained to fill with a textured pattern; b) the action of filling with a textured pattern, even when applied to a novel object.

A more complex example comes from another participant, who makes an interesting set of generalisations of the label 'suqimogi' (Figure 6.9). The way this label is generalised suggests a pattern like the example above, where object labels derive from their habitual actions. However, in this case, the label is also extended to label the action itself performed by a novel object. Importantly, this label is neither an unambiguous object label or action label, but its pattern of reuse is still systematic. This shows an experimental example of the kind of chaining-derived polysemy discussed in Chapter 2. A parallel from natural language is the polysemy exhibited by the word 'cook', which can refer to an action or to a person who habitually performs this action.

This pattern was produced by only one participant in the pair. Indeed, participants within a pair frequently generalised their communicative language in notably different ways during the post-test. It should be emphasised that the general patterns above emerge out of a great deal of variation, both within- and between-pair, in the features lexicalised and the strength of these associations over the phases of the experiment.

6.6.5 Problems and future directions

6.6.5.1 Stimuli

The co-occurrence structure of the objects and actions in the experiment combines many interdependent constraints in a potentially non-intuitive way. This could have confused participants and hampered their mapping of the structures to patterns of word reuse. A potentially more ecologically valid manipulation for a future experiment would be to use events involving an agent, an action, and a patient. In this setup, the same objects could act as both agents and patients, with their predictability in different roles varying systematically as a function of particular actions. This would be more similar to the constraints on real-world events discussed in the Introduction, and could be more accessible for participants, while still possible to model in an ‘alien’ world with an artificial language to minimise native language interference.

6.6.5.2 Event test: inferring possibility

As noted in the Results and Discussion, the task of judging whether given events were possible or impossible may have confused participants. During the experiment, participants are given only negative evidence for the ‘impossibility’ of particular object/action/action combinations; this may not be sufficient to provide a strong expectation that these events are truly impossible. A non-binary way of measuring participants’ acquisition of event structure constraints would be more informative. One possibility is a task more like the dimension selection task from Voiklis & Corter (2012), as used in Experiment 1: participants could be shown an object or an action and asked to predict the remaining features of the event. The distribution of their responses would give a more nuanced view of their event structure expectations than a binary possible/impossible judgement, as well as tying in more directly to the hypothesis that it is features’ predictive value that determines their likelihood of lexicalisation.

6.6.5.3 Post-test: participant fatigue and frequency differences

Many participants did not enter a label on many trials on the post-test. The option of not entering a label was deliberately left open for participants, in order to capture event features for which they did not have a label: e.g., if participants do not type a label for a constrained action involving a novel object, we can infer that they do not have an independent label for this action. However, it is possible that by this

stage of the experiment participants were simply skipping trials in order to finish the experiment faster. The post-test may therefore underestimate the lexicalisation of all features. The lack of a significant main effect of experimental phase on association strengths suggests that participants still provided enough data to be confident about the associations found by the analysis. However, future work could require a more active response from participants in cases where they did not have a label for a given trial, rather than simply pressing Enter.

An additional potential concern is that, while frequencies of constrained vs. unconstrained objects and constrained vs. unconstrained actions were balanced in the communication phase, this was not the case for the post-test. For example, in the post-test, each unconstrained action appeared more frequently than each constrained action. This greater frequency may have made these actions more salient to participants, leading them to lexicalise them more strongly (although, correspondingly, the ‘burden of proof’ for lexicalisation of unconstrained actions is greater, since all association values are calculated as proportions of total meaning feature occurrences). In future, a more balanced design could help to address this potential problem. However, total balance will be a challenge to achieve while still maintaining the manipulation of constrained co-occurrence.

6.6.5.4 Task demands in communication vs. post-test

The communication task in the experiment was extremely challenging for participants. On the basis of just 2 words of training, they had to a) infer a segmentation for the labels they had learned, b) extend parts of these labels to novel events, c) invent labels for event features they had not seen, and d) coordinate on conventions with their partner. The lack of a significant difference between constrained and unconstrained actions in the communication phase could be a result of this very demanding task. In addition, the association data from the communication phase are collapsed over all trials; however, during the communication phase, participants’ languages were constantly changing, making early data points potentially misleading. Future versions of the experiment could run the communication phase for longer and only collect association data from later rounds, as for the category systems in the experiments presented in Chapter 5. Notably, the original Pirog Revill et al. (2008) study used a two-day training program for participants, even though they had the easier task of learning a lexicon, not generalising one from minimal training. Giving participants more time to consolidate their knowledge could also potentially lead to stronger effects emerging in

the communication phase.

In addition, the format of the post-test may itself exert different pressures on participants. The format of a labelling task, where each event is segmented with its parts presented in isolation, could activate meta-linguistic knowledge (for example, pushing participants to conceptualise actions as requiring verb labels). Given that every event and combination is presented, this phase may also have created the impression that the experimenter wanted a label on every trial, creating an artificial pressure for generalisation. It is important to note that the main result still stands: any general expectation of this kind would not account for the difference between constrained and unconstrained actions in the post-test phase. However, more generally, the role of metalinguistic knowledge in the kind of adaptive generalisation we see in the post-test could be investigated further, for example by running the communication phase for longer and seeing whether these adaptive generalisations emerge in a less metalinguistically charged context.

6.7 Conclusion

Language comprehension is a process of incremental prediction of utterance meaning from the interaction of world knowledge and linguistic conventions. The experiment presented in this chapter tested the hypothesis that meaning elements that are less predictable across contexts will tend to be lexicalised more strongly in an emerging language.

As in Experiments 3 and 4, communication without prior learning of a structured language is noisy, leading to reliance on more salient meaning features (in this case, objects) for the development of initial conventions. However, subsequent individual generalisation from these noisy data shows patterns that conform to the hypothesis: patterns of word reuse tend to be associated more strongly with meaning features that are overall less predictable, given the structure of the world and the other conventions in the language. In an individual generalisation task, where the language was no longer being used for communication, participants inferred a language that was adapted to be expressive given the structure of events in the world of the experiment. The results show that patterns of event structure in the world, communicative inference, and individual learning and generalisation all interact to shape the patterns of word reuse in an emerging language.

CHAPTER 7

Conclusion

This thesis began with the observation that language gives humans the unique ability to communicate an open-ended range of meanings. The key to this ability is the apparent correspondence of units of language (broadly characterisable as words) to units of meaning, such that we can combine words to communicate complex meanings. Understanding the origins, structure, and cognitive representation of word meanings is therefore crucial for understanding how language evolved.

Word meanings have often been argued to correspond to stable, discrete concepts that exist prior to and independently of language. The argument of this thesis has been that this picture, despite being intuitively appealing and parsimonious, is misleading. Instead, word meanings emerge as a by-product of coordinating on and culturally transmitting communicative conventions. Words are learned from their contexts of use and reused according to generalisations made on the basis of features that are more salient in linguistic and non-linguistic contexts. The flexible and emergent nature of these generalisations enables the continual establishment of new uses for existing conventions, allowing language to remain both expressive and learnable.

Evolutionarily, word meanings can be characterised as the result of an evolved capacity for inferring the intentions of others from their behaviour and from known patterns of events in the world. The establishment of intentional communication allows conventions to be coordinated on and passed on via cultural transmission. Under this view, the structure of word meanings is shaped by two main pressures: 1) learnability, or the pressure for word meanings to be acquirable from observation of communicative

contexts; 2) expressivity, or the pressure for word meanings to contribute efficiently to the inference of utterance meaning during communication. Both of these pressures interact with the structure of the world as filtered through human perception (including properties of objects and actions, and typical event structures in which they combine).

The experiments in the thesis tested the effects of these pressures on the structure of word meanings. Experiment 1 found that low-level individual learning effects amplified by cultural transmission led to strong patterns of underspecification on dimensions that were habitually less salient in learning and production contexts. Broadly, this shows that adaptive patterns of underspecification that reflect the salience of particular dimensions can emerge gradually as a product of iterated learning, and need not be pre-specified in non-linguistic conceptual structure. This supports the theoretical model of meaning presented Chapters 2 and 3 by showing that the dimensions of the meaning space that are relevant for labelling change as a consequence of participants' patterns of word reuse, rather than reflecting the relevance of dimensions for labelling in the language they learned. Crucially, this outcome does not result from intrinsic salience of particular dimensions of the referents, but salience in contexts of learning and use. This supports an account of word meanings as contexts of use, where the dimensions highlighted or backgrounded by these contexts influence future patterns of reuse. These systematic contextual effects work together with learning pressures to push participants to reuse words along contextually salient dimensions, changing the overall dimensional structure of word meanings in the language.

In Experiment 3, coordinating on communicative conventions for images drawn from a continuous space led to word meanings that were sub-optimal for communication, compared to word meanings produced by non-communicating individuals. Communicating pairs' word meanings were also less aligned than those of individuals who did not communicate. Experiment 4 found that when cultural transmission was added to this picture, word meanings became more optimally structured for communication and more aligned within pairs over generations. In addition, word meanings became more aligned across pairs over generations, although within-pair alignment remained consistently higher. This suggests some level of convergence to universal structures, partly explained by a shared similarity space (as described in Experiment 2), but with some lineage-specific effects stemming from particular transmission and interaction histories. Alignment tended to be localised to perceptually salient regions of the stimulus space. This fits well with the mix of universality and diversity we find in word meanings across languages, where words in different languages may share prototypi-

cal (more frequent or more experientially salient) uses, but diverge on more peripheral uses. More broadly, the results suggest that both learning and communication are required for the emergence of structured, aligned word meanings. Learning reduces the number of meanings, allowing the system to remain stable over cultural transmission, and communication increases the structure and alignment of word meanings within a speech community on the basis of this learned common ground. Relating this back to the theoretical model, this supports an account of word meanings as patterns of word reuse that are optimised for communication given the structure of the world, as well as being aligned within communicating pairs. However, surprisingly, this is not the case for improvised communication, but only where communication is preceded by learning. This supports the hypothesis advanced in Chapter 3 of word meanings as a culturally evolved compromise between learnability and expressivity. Furthermore, the higher within-pair than across-pair alignment specifically in the communicative conditions provides an experimental insight into the establishment of different communicative extensions of words across languages, as in the examples from Greek and English presented in Chapter 2.

Experiment 5 found that constraints on the co-occurrence of objects and actions influenced the event features that are lexicalised in a developing language. During communication, constraints on possible events did not have a strong effect on patterns of word reuse; however, in a post-test where participants were asked to generalise the language they had used during communication, event features that were less predictable from linguistic and non-linguistic context were lexicalised more strongly. This suggests that communicators' common ground knowledge of event structure has an effect on their patterns of word reuse, but not necessarily during communication: rather, it has an effect on the patterns they infer when asked to generalise. More broadly, in conjunction with the results from Experiment 3, the results reveal the noisiness of communication without a previously learned system, and suggest that the generalisation and rationalisation that comes from individual learning is an important stabilising influence on word meanings. The results support the theoretical account of words as optimally reused cues for utterance interpretation in the light of known patterns of event structure in the world; however, while inferential communication was hypothesised as the mechanism for this optimal reuse, the effect in the experiment only appeared in a post-communicative generalisation test that may have tapped into specifically metalinguistic knowledge. The discussion in Chapter 6 suggests follow-up studies that could investigate this further.

Taken together, these results imply a different evolutionary picture of word meanings from the one presented at the beginning of Chapter 3. Rather than being discrete chunks of conceptual structure that exist prior to language, word meanings are shaped by the fact that they must be learned from and used in communicative contexts. Pre-linguistic conceptualisation is not the only determinant of word meaning structure: instead, it is one of a number of influences, including pressures specific to language as a system of culturally transmitted communicative conventions. The consequence is that the systematicity of word meanings is only partial. Theories that argue for complete systematicity, e.g. via lexical decomposition into discrete conceptual primitives, may be more of a descriptive abstraction than an accurate characterisation of how word meanings are learned and mentally represented. Word meanings are only as systematic as a) the world and b) human generalisation over communicative inferences, understood as a context-, task- and history-dependent process.

This thesis has presented preliminary work on some specific questions about the influences of learning and communication on word meanings. Many questions still remain to be investigated. The experiments use simple perceptual models of world structure: discrete feature-based stimuli in Experiments 1 and 5, and continuous variation between generated exemplars in Experiments 2, 3 and 4. The structure of the real world is richer and more complex. As discussed in Chapter 3, categories in the real world have deep covariational structure: the possession of particular features is correlated with the possession of others, and these rich correlations provide more of a basis for our categorisation and word reuse than is available from the relatively impoverished stimuli in the experiments. In addition, we do not categorise or communicate about objects and actions based on their perceptual similarity alone, but in the service of particular goals, for which function has equal or greater importance than perception. The intention of using perceptual similarity in the experiments was to control for function as a further variable that might influence patterns of word reuse. However, in the real world, perceptual similarity and function are interdependent; for example, we can tell a stick will make a good weapon if it looks pointy. Furthermore, in relation to the model of word meaning advanced in Chapter 2, perceptual similarity and function are context-dependent. For example, different implied functions for perceptually identical objects lead participants to weight features differently for categorisation (Lin & Murphy, 1997). Future experimental work could actively investigate these complexities rather than abstracting away from them. For example, an experiment where the same objects are used for different functions, which in turn highlight differing perceptual fea-

tures – such as using blocks for building towers versus bridges – could illuminate how contextual effects change patterns of word reuse and their relation to perceptual and functional features. More generally, varying the extent to which the stimuli are structured on perceptually obvious dimensions could address whether a relative absence of structure in the world leads to a greater influence of learning and communication on word meaning structure. A followup question of interest is the extent to which these structures would tend to be universal and optimal, or constrained and lineage-specific. Another possible approach would be to use stimuli with a mix of concrete perceptual dimensions and more abstract dimensions (for example, displacement in time). This would allow investigation of how conventions established for features on concrete dimensions are then generalised to more abstract uses, as in real language where spatial terms are generalised to the domain of time (Lakoff & Johnson, 1980).

The effects of learning could also be investigated further. In the experiments presented in this thesis, there is no learning bottleneck – i.e., selection of a subset of word-meaning pairs for learners to be exposed to. Given the importance of the bottleneck to models of the emergence of linguistic structure (Smith et al., 2003a), the effect of this variable should be investigated. However, in the model of word meaning used in this thesis, the possible uses of a word do not form a finite set, since words can always be flexibly generalised to new uses. Therefore, rather than the presence or absence of a bottleneck, the key variable is how many uses of a word the learner is asked to generalise from, and how this interacts with the number of uses they can remember (i.e., a memory-based bottleneck internal to the individual learner). The contexts of word use that a learner is exposed to may influence their generalisations in a number of ways. For example, being exposed to only few uses may constitute more ambiguous evidence for the appropriate contexts of use of a word, causing different learners to generalise along different dimensions. A word's contexts of use can also be more or less diverse, i.e., share more or fewer features. Training a learner on contexts that are more or less diverse may influence their patterns of reuse, leading the learner to condition their use of a word on more or fewer features (Landau & Shipley, 2001; Xu & Tenenbaum, 2007), with potential knock-on effects for the higher-order generalisations learners make about the dimensions on which the reuse of words should be conditioned. More broadly, it should be borne in mind that participants' already-learned biases about the appropriate dimensions on which to reuse words (such as the shape bias) will influence their responses in artificial language experiments. While the studies presented in the thesis control for this by shuffling or counterbalancing the dimensions of the

stimuli (Experiments 1 and 5) or using stimuli with unfamiliar dimensional structure (Experiments 2-4), future work could use artificial language learning experiments with child participants to actively investigate the processes by which these biases become established. A precedent for work in this area is Smith et al. (2002), where a 9-week longitudinal study with 17-month-old children found a shape bias emerging gradually over the course of the experiment for children who were trained on categories organised by shape. Future work could investigate whether different biases, including those that are not necessarily reflected in a child's target language, can be induced by similar periods of training in this age group.

The model of communication used in the experiments is also simplified: for example, all communication takes place in pairs. While dyadic communication can serve as a simple model for communication in general, increasing the number of participants in a communicative interaction could create a more complicated picture. If coordinating on conventions in a pair is a noisy process, as shown in Experiments 2 and 4, then the addition of further participants could slow this down further, and even hold back the communicative alignment after learning that occurs in Experiment 3. Another simplification is that the communication games involve unambiguously presented targets for the sender, and a choice of discrete targets for the receiver. Future work could use more task-based experiments, where meanings arise as a by-product of a communicative task requiring the flexible reuse of conventions. For example, a task which can be solved by generalising words on any of a number of dimensions (e.g., by conventionalising either colour terms or shape terms) would allow further investigation of communication as an instigator of shared, adaptive ways of abstracting across groups of stimuli. While Experiment 1 of this thesis showed that iterated learning can create patterns of adaptive abstraction based on dimensions that are strongly cued as salient, an experiment like the one described above could answer the question of whether communication creates, and/or encourages interlocutors to align on, particular abstractions where several are possible. This could provide a clearer experimental analogue for the chaining processes that lead to cross-linguistic variation in word meaning extensions, as discussed in Chapter 2.

Another aspect of communication not explored extensively in this thesis is the receiver's side of the interaction. Aside from the brief discussion of receivers' preferred exemplars in Chapter 5, word meanings have been described in terms of patterns of word reuse by the speaker. However, since the listener's inferences when they hear a word will influence the contexts in which they choose to reuse it and hence the evol-

ing word meaning in the speech community, this is an aspect of communication that deserves further investigation. Alternative experimental methods – for example, a communication game where each trial involves not a binary correct/incorrect answer, but selection of a range of possible referents – could allow investigation of how different contexts affect the range of possibilities the receiver is willing to infer. Likewise, as discussed in Chapter 5, different models of feedback could incentivise differing word meaning structures. Less fine-grained feedback could promote correspondingly less fine-grained word meanings, and could also illuminate the extent to which alignment is contingent on particular communicative tasks, rather than necessary for communication in general to succeed.

More generally, it could be argued that the separation of learning and communication is somewhat artificial. Rather than being two discrete stages in the development of word meanings, they are intertwined. This interdependence goes both ways: not only are word meanings learned from communicative contexts, but the use of a word in a communicative context relies on the speaker's learned exemplars of its previous uses, whether this learning precedes communication (as in Experiment 4) or happens during communicative negotiation (as in Experiment 3). Future work could focus on this interdependence. For example, gradual turnover experimental designs, where naive participants learn from observation of and participation in communication before being replaced by new naive participants, could provide a model that more closely mirrors the inseparability of learning and communication in the acquisition and use of real languages.

Learning and communication are not the only pressures at work in shaping word meanings. A pressure alluded to in Chapter 2 but not investigated in this thesis is frequency: the extent to which particular objects, actions or events crop up more often than others in the meanings of utterances. This is not the same as frequency or typicality of objects or events in the world. An event can occur very frequently without being frequently communicated about: for example, we do not generally remark on how much we are breathing. Future experiments could manipulate the frequency of communication about particular events independent of their frequency of occurrence, investigating whether communicative frequency trumps frequency of occurrence in determining which event features are lexicalised and/or the dimensions on which words are generalised.

A final issue not addressed in this thesis is the extent to which word forms may directly reflect their meanings. While not investigated here, sound-symbolic or other

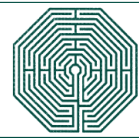
iconic correspondences between word and meaning are likely to have an impact on both the learnability of word meanings (Monaghan et al., 2012; Lupyan & Casasanto, 2014) and the inferrability of meaning in communicative contexts (Garrod et al., 2007). In particular, sound-meaning or gesture-meaning correspondences could help to solve the problem of the origins of the first communicative conventions. To elaborate: the account presented in this thesis is that a word is just one clue to meaning, working with non-linguistic and linguistic context to constrain the possible meanings of an utterance. This implies that before linguistic conventions were established, non-linguistic context would have been the only evidence available for inferring the meaning of novel conventions. However, if non-linguistic context was sufficient for inferring meaning, then communication would be redundant, offering no motivation for establishing a convention at all. A way out of this paradox is an initial intrinsic connection between signal and meaning in the form of iconicity. Once some conventions are established, these may then form a linguistic context that interacts with the world knowledge and inferential abilities of interlocutors to allow for the establishment of novel conventions, without requiring an intrinsic connection between form and meaning (Ramachandran & Hubbard, 2001; Perniss et al., 2010; Cuskley & Kirby, 2013). However, this account is speculative and requires further empirical investigation.

To summarise: word meanings have traditionally been modelled as discrete chunks of conceptual structure that exist prior to language. I argue for a different picture, where word meanings are conventions, learned from their contexts of communicative use, stored as exemplars of these contexts embedded in a world knowledge network, and generalised on the basis of features of these exemplars. This provides an alternative account of how language can be flexibly expressive while consisting of a limited number of learned conventions. The implications of this view are that word meanings are shaped not only by the structure of the world as filtered through human perception, but are also by processes of iterated learning and communicative use. In this thesis I presented a series of experiments demonstrating the effects of these pressures on the structure of word meanings. The results open up new experimental methods for investigating the origins and development of word meanings, as well as giving insights into their enabling role in the unique communication system that is human language.

APPENDIX **A**

Published Papers

Reproduced in this section are Silvey et al. (2014) and Silvey et al. (2013).



Word Meanings Evolve to Selectively Preserve Distinctions on Salient Dimensions

Catriona Silvey, Simon Kirby, Kenny Smith

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

Received 21 August 2012; received in revised form 13 December 2013; accepted 27 December 2013

Abstract

Words refer to objects in the world, but this correspondence is not one-to-one: Each word has a range of referents that share features on some dimensions but differ on others. This property of language is called underspecification. Parts of the lexicon have characteristic patterns of underspecification; for example, artifact nouns tend to specify shape, but not color, whereas substance nouns specify material but not shape. These regularities in the lexicon enable learners to generalize new words appropriately. How does the lexicon come to have these helpful regularities? We test the hypothesis that systematic backgrounding of some dimensions during learning and use causes language to gradually change, over repeated episodes of transmission, to produce a lexicon with strong patterns of underspecification across these less salient dimensions. This offers a cultural evolutionary mechanism linking individual word learning and generalization to the origin of regularities in the lexicon that help learners generalize words appropriately.

Keywords: Attentional learning; Cultural transmission; Iterated learning; Language evolution; Word meaning; Language and conceptualization

1. Introduction

Language allows us to communicate about the world. This is possible because parts of language (e.g., words) refer to parts of the world (e.g., objects). However, this relationship is rarely one-to-one. For example, the word “cat” refers to a range of objects that share features on certain dimensions, such as shape, but differ on others, such as color. This abstraction over features is a ubiquitous property of natural language called underspecification (Geeraerts, 2009, p. 196).

Different areas of the lexicon have different characteristic patterns of underspecification. For example, words for artifacts tend to specify shape or function, and underspecify

color; words for substances tend to specify material, and underspecify shape (Smith & Samuelson, 2006). These regularities in the lexicon enable learners to acquire higher order generalizations about which dimensions are relevant to the meaning of words learned in particular contexts, for example, the shape bias that labels for objects generalize by shape (Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002).

However, this account does not explain how the lexicon comes to have these helpful regularities in the first place. One possibility is that learners have strong constraints on the kind of word meanings they will entertain (Markman, 1994; Waxman & Kosowski, 1990), which map straightforwardly to strong constraints on the kinds of underspecification lexicons can exhibit. Instead, we show that the same processes that enable learners to form higher order generalizations on the basis of regularities in the lexicon can also shape the lexicon to exhibit those regularities in the first place, leading it to reflect the systematic salience of particular dimensions in contexts of learning and use. This happens not over the course of an individual's learning, but via the cumulative language change that occurs when a lexicon is transmitted.

The attentional learning account states that “context cues that co-occur with (and define) specific tasks will come with repeated experience to shift attention to the task-relevant information” (Smith, Colunga, & Yoshida, 2010, p. 1295). Modeling the learning of (part of) the lexicon as this kind of “specific task,” we train and test learners on an artificial language in contexts where one dimension of meaning is systematically made less salient (backgrounded). We manipulate salience by casting word learning and use as a series of discrimination games where one dimension is never helpful. The general format of the discrimination game has a precedent in the “guessing game” of Steels (2003), while manipulating one dimension to be unhelpful builds on the well-established results in the concepts and categories literature showing that dimensions that are unhelpful for discrimination are attended to less than helpful dimensions (e.g., Kruschke, 1992; Medin & Schaffer, 1978). In real word learning, this backgrounding effect is more likely the outcome of factors such as domain-specific knowledge (Kelemen & Bloom, 1994; Lin & Murphy, 1997), increased salience of functional features (Booth & Waxman, 2002; Keil, 1994; Kemler Nelson, 1995), attentional cues from speakers (Tomasello, 2000), inference of the speaker's intention (Bloom, 2000; Xu & Tenenbaum, 2007), or other “non-linguistic evidence of the speaker's locus of attention” (Clark, 1997).¹ This systematic backgrounding has only a small effect at the individual level. However, over cultural transmission, a lexicon that initially specifies equally across all dimensions changes to reflect the differing salience of dimensions in learning and use, leading to an emerging system which preferentially underspecifies the backgrounded dimension. This serves as a demonstration of how cultural transmission amplifies the effects of individual learning processes to create an adaptively specified lexicon.

1.1. Modeling the cultural evolution of underspecification: Iterated learning

We model the cultural evolution of language using iterated artificial language learning (Kirby, Cornish, & Smith, 2008; Smith & Wonnacott, 2010). In the diffusion chain

instantiation of this paradigm, participants are organized into chains of transmission: An initial language is taught to the first learner in each chain, who subsequently attempts to reproduce that language; this reproduction is then given as learning input to the next participant in the chain, and so on. Using this methodology, researchers have demonstrated the cultural emergence of properties of language including arbitrariness (Caldwell & Smith, 2012; Fay, Garrod, Roberts, & Swoboda, 2010; Theisen-White, Kirby, & Oberlander, 2011), regularity (Reali & Griffiths, 2009; Smith & Wonnacott, 2010), categorization that reflects discontinuities in world structure (Perfors & Navarro, 2011), compositional structure (Kirby et al., 2008, Exp 2; Theisen-White et al., 2011), and underspecification (Kirby et al., 2008, Exp 1).

Our method here is based on Exp 1 from Kirby et al. (2008). The “meanings” in this study were a series of images that varied in shape (square, circle, triangle), color (black, blue, red), and motion (horizontal, bouncing, spiraling). Each chain was initialized with a language which provided a unique word for each of the 27 meanings; that is, it specified fully across all dimensions. However, due to the difficulty of accurately learning and reproducing this language given the amount of training provided, participants began to reuse words for referents that differed on certain dimensions. This led, over several generations of transmission, to the emergence of underspecification as a solution to the learning problem; for example, in one chain, every bouncing square came to be labeled “tupim,” regardless of color.

However, this underspecification was not consistently directed to any particular dimension. Across the different chains, some languages underspecified color, some shape, and some motion (Cornish, 2011), presumably because, in the learning and testing procedures used in Kirby et al. (2008), no particular dimension was made more or less salient. In contrast, in real word learning and use, some dimensions have higher salience than others (Clark, 1993; Regier, 2005). For particular groups of referents, commonalities across these situations of learning and use will result in certain dimensions being foregrounded and others backgrounded, as per the attentional learning account (Smith et al., 2010). Our hypothesis is that these systematic differences in dimension salience during individuals’ learning and production can lead, over cultural transmission, to a pattern of underspecification that reflects these differences—a helpful lexicon that aids subsequent learners in making the right kinds of generalizations. To test this hypothesis, we ran a modified version of the Kirby et al. (2008) paradigm, where the learning and production procedures are structured to systematically background one meaning dimension: Meanings are presented in pairs that share a feature on one consistent dimension, such that attending to this dimension will never help participants discriminate between the two meanings (Fig. 1). The hypothesis is that underspecification will gradually arise on the backgrounded dimension, thus showing that strong constraints on learners’ word meaning hypotheses are not necessary to explain the patterns of underspecification we see in natural language. If, on the other hand, underspecification were to arise indiscriminately on all dimensions (as in Kirby et al., 2008), this would suggest that stronger constraints are needed to explain real-world patterns.

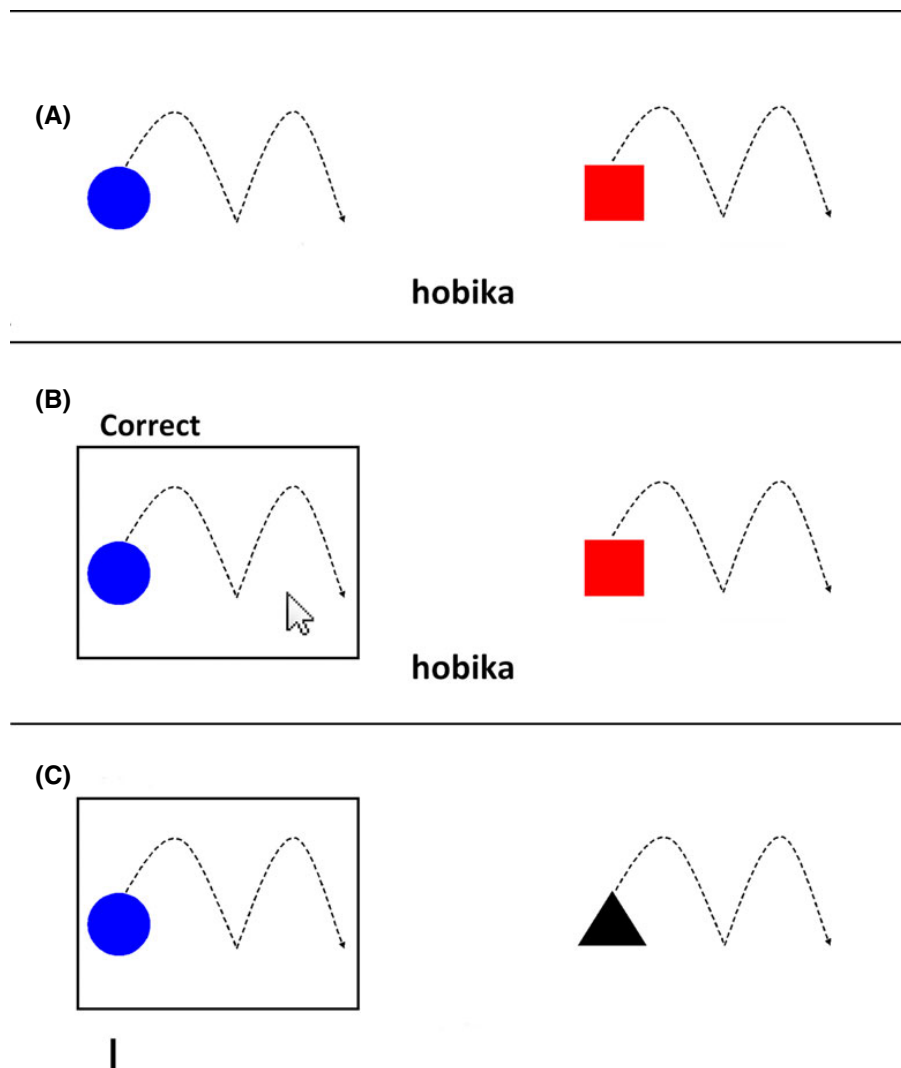


Fig. 1. Training and testing procedures in the experiment. (A) Each training trial is presented as a discrimination game. The participant is shown a word and two candidate images. The participant clicks the image he or she thinks goes with the word. (B) The participant is then given feedback, followed by the correct word-image pairing. The word then disappears and they are required to retype it. (C) Test trials are again presented as a discrimination game, but from the opposite perspective. The participant is presented with two images, one of which is selected as the target. The participant is instructed to type the word that would allow the alien to pick the correct image. In all training and test trials, target and distractor share a feature on one consistent dimension (in this example, the motion dimension).

2. Method

2.1. Participants

Forty undergraduate and graduate students at the University of Edinburgh (25 female, median age 20.5) were recruited via mailing lists and organized into eight diffusion chains. Each chain consisted of an initial participant who was trained on a random language, and four successive participants who were trained on the previous participant's test

output language, making five generations in total; the results of Kirby et al. (2008) suggest that five generations would be sufficient for underspecification to arise (in three of their four chains the languages had fewer than five words by generation 5). Participants in chains 1–6 were unpaid volunteers; participants in chains 7–8 were paid £4.50.²

2.2. *Stimuli: Images and input language*

Participants were asked to learn and then produce an “alien language,” consisting of lowercase text labels paired with images. The images were the 27 pictures of colored shapes in motion from Kirby et al. (2008). The images varied in three possible ways on each of three dimensions of color, shape, and motion (see Fig. 1 for examples). The training language for the first participant in each chain was a randomly generated set of 27 unique two- to four-syllable labels, built up from nine possible CV syllables (“da,” “vi,” “ho,” “wi,” “nu,” “ri,” “bi,” “ka,” “tu”). These labels were randomly assigned to the 27 images, ensuring that there was a unique label for every image, with no systematic structure to the labels. The training language for later participants was the language produced by the previous participant in the chain during testing.

2.3. *Procedure*

2.3.1. *Language learning, language testing, and dimension selection task*

The participants worked through a computer program with three phases:

1. Learning phase.

In each learning trial, the participant was presented with a label and two images, one of which was the target and one a distractor. The participant was instructed to pick which of the two images corresponded to the label. Once the participant had clicked an image he or she was told whether the choice was correct or incorrect, shown the label and correct image for 2 s, and then instructed to retype the label before proceeding to the next trial. Target images were presented in random order. Distractors for each trial of the learning phase were assigned at random, subject to the following constraints: (a) within each learning block, each of the 27 meanings appeared once as a target and once as a distractor; (b) according to the main experimental manipulation, one dimension was consistently backgrounded during learning and testing trials. For each participant, one of the three dimensions of shape, color, and motion was selected as the backgrounded dimension. Every distractor then had the same feature as the target on this dimension (for example, if color was selected as the backgrounded dimension, the distractor on every trial would be the same color as the target). The other two dimensions were not manipulated in this way and served as controls. The learning phase of the experiment consisted of four blocks, each of 27 trials.

2. Test phase.

In each test trial, the participant was presented with two images: a target and a distractor. The target was highlighted with a black border. The participant was instructed to type the

label that would let the alien know which image was highlighted. Target images were presented in random order. Distractors were randomly assigned within the same constraints as in the learning phase; that is, they matched the target on the backgrounded dimension. The test phase consisted of 27 trials, one for each target.

3. Dimension selection task.

This final phase of the experiment used a method from Voiklis and Corter (2012) to test which dimensions participants thought essential to word meaning. On each trial, participants were presented with a label from the language they had been trained on and a concealed image. Their task was to decide whether the label-image pairing was correct or incorrect. To do this, they could click to reveal a feature of the concealed image (shape, color, motion), in any order. Participants could click correct or incorrect at any stage and did not have to reveal all features before doing so. A 1 s delay was included before features were revealed, to discourage participants from revealing features which were unnecessary to make the correct/incorrect judgment. The dimension selection task consisted of 27 trials, one for each image. Images were presented in random order. The labels for each trial were selected from the language the participant was trained on, such that 14 trials contained correct picture-label pairings and 13 incorrect picture-label pairings, but each label appeared only once.

2.3.2. *Iteration*

The language each participant produced in the test phase of the experiment was transformed and then used as the training language for the next participant in his or her chain. For this transformation, all dimensions and features of the images were randomly shuffled, so that patterns of labeling in relation to backgrounded and control dimensions were preserved, but individual correspondences of labels to images were not (see Fig. 2 for an example). This transformation was intended to reduce the effects of intrinsic differences in salience of different dimensions, and to prevent the establishment of iconic labels (e.g., reduplicated syllables for bouncing images).

2.4. *Dependent variables*

We used the measure of transmission error (how much the language produced by a participant during testing differed from their training input) from Kirby et al. (2008) to test whether the languages became more learnable over generations. Normalized Levenshtein edit distance between corresponding labels in successive generations (e.g., “taho” and “takiwi” for meaning 3 in Fig. 2) was 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. These values were then averaged across the whole language to give one measure of error per participant. If this value decreases over generations, the language is becoming more learnable.

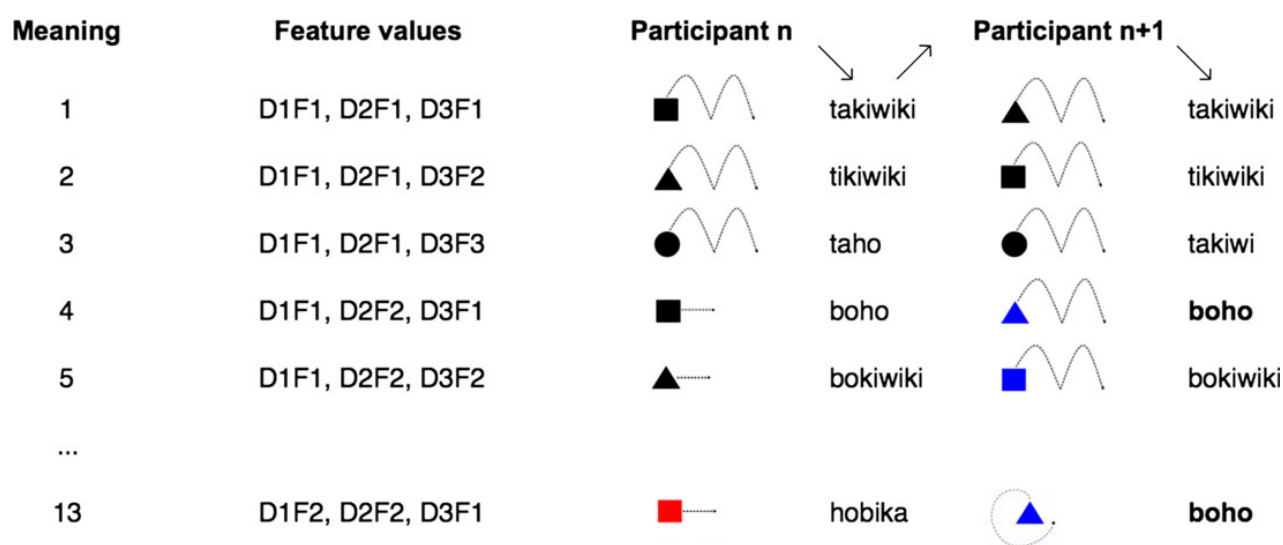


Fig. 2. Illustration of the transformation process between participants in a chain. In the rest of this legend, “bk,” black; “rd,” red; “bl,” blue; “ci,” circle; “sq,” square; “tr,” triangle; “ho,” moving horizontally; “bo,” bouncing; “sp,” spiraling. During the test phase (downward arrow), participant n produces mappings between 27 meanings (obtained from three features on Dimension D1 \times 3 features on Dimension D2 \times 3 features on Dimension D3) and 27 labels. The meaning of each label is therefore represented by specifying the Dimensions D1, D2, and D3 with features F1, F2, and F3 for each dimension. For example, for participant n , D1 is Color (where F1 = bk, F2 = rd, F3 = bl), D2 is Motion (where F1 = bo, F2 = ho, F3 = sp), and D3 is Shape (where F1 = sq, F2 = tr, F3 = ci). D1 is the backgrounded dimension (here, Color). The labels produced by participant n are presented to participant $n + 1$ during the training phase (upward arrow); however, their corresponding meanings (i.e., the pictures) are changed randomly (while preserving the backgrounded and control dimensions). In the example, for participant $n + 1$, D1 is Motion (where F1 = bo, F2 = sp, F3 = ho), D2 is Color (where F1 = bk, F2 = bl, F3 = rd), and D3 is Shape (where F1 = tr, F2 = sq, F3 = ci). The backgrounded dimension is still D1, but it is now Motion rather than Color. The final column shows the new label produced by participant $n + 1$ during their test phase (downward arrow). Here, we can see that while for participant n “boho” means “black square moving horizontally,” and “hobika” means “red square moving horizontally,” for participant $n + 1$ “boho” means “blue triangle” regardless of motion. In other words, for meaning 13, this participant produces “boho” where they were trained on “hobika,” changing the language with this error to introduce underspecification across the backgrounded dimension (motion for this participant).

Our specific hypothesis was that the languages would evolve gradually to underspecify more on the backgrounded dimension than on the control dimensions. Three outcome measures were taken to assess whether this was happening.

To capture the extent to which a language made distinctions on each dimension, we calculated (for each participant’s test output): (a) the average number of words the language used across the features on each dimension (possible values ranging between 1 and 3); (b) average normalized Levenshtein edit distance between these labels, to give a more fine-grained measure of label dissimilarity. Fig. 3 gives an example of how these measures were calculated.

Thirdly, participants’ behavior on the dimension selection task (the order in which they chose to reveal the dimensions) was used to quantify participants’ attention to particular dimensions when evaluating word meaning. We gave a score of 3 for the dimension clicked first, 2 for second, 1 for third, and 0 if the dimension was not selected at all.










Meaning group	Number of words	Word dissimilarity
 nununu nununu nununu	1	0
 rutunu rutunu rutunu	1	0
 nununu ruhovu hovani	3	0.78
 rinunu rinununu rinununu	2	0.17
 rinunu rutunu rutunu	2	0.22
 rihonu rihova rihova	2	0.22
 danunu danunu danunu	1	0
 danunu danunu danununu	2	0.17
 dahovu dahovu dahovu	1	0
	1.67	0.17

Fig. 3. Sets of meanings whose labels were compared to obtain the measures of language structure (here, with respect to the motion dimension). Meanings were divided into sets of three that differed only on one dimension. The word dissimilarity score is calculated by averaging the three normalized Levenshtein edit distances obtained by comparing the three possible word pairs. For example, for row 4, rinunu/rinununu = distance 0.25, rinunu/rinununu = distance 0.25, rinunu/rinunu = distance 0, so the average word dissimilarity is 0.17. (Normalized Levenshtein edit distance for rinunu/rinununu: two letter additions necessary to turn one word into the other, divided by the length of the longest word (8) = 0.25.) Similar measurements are then made over all nine sets of three meanings that differ only on the motion dimension, and these values averaged to give one underspecification value for that dimension for number of words, and one for word dissimilarity (values in bold).

Dimensions which are selected earlier, and are therefore presumably more central to word meaning, will have higher scores.

3. Results

Transmission error consistently decreased over generations ($M_{\text{generation } 1} = 0.67$, $SD = 0.10$; $M_{\text{generation } 5} = 0.34$, $SD = 0.16$). A linear trend ANOVA found that the trend was significant, $F(1, 7) = 27.84$, $p < .001$, showing that the languages changed to become more learnable.

The results for the edit distance measure of underspecification are shown in Fig. 4, with the result for the dimension selection task in Fig. 5. Mixed-effects models were used for the main analyses of each of our dependent variables (number of words, within-dimension label dissimilarity, dimension selection task).³ p -values for the fixed effects in these models were estimated using Baayen (2008)'s formula.⁴ For post-hoc tests, the observations for the two control dimensions were averaged. Between-group t -tests were then run comparing backgrounded and control dimensions at each generation, applying the Bonferroni correction for multiple comparisons.

1. Number of words. Mean number of words across backgrounded and control dimensions was similar at generation 1 ($M_{\text{backgrounded}} = 2.93$, $SD = 0.10$; $M_{\text{control}} = 2.91$,

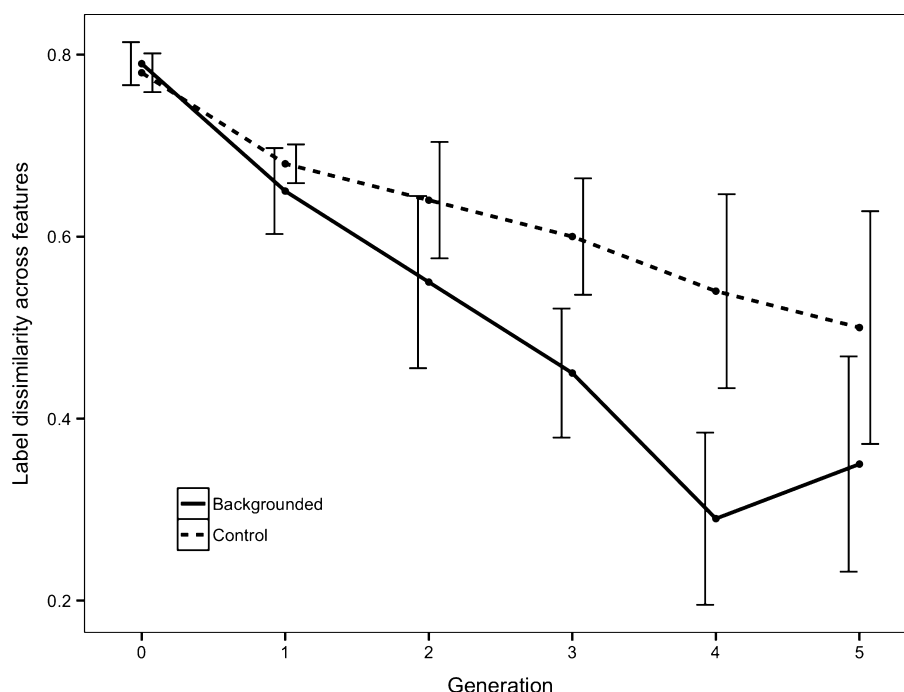


Fig. 4. Dissimilarity of labels across features (see Fig. 3 for how this is calculated) against generation. The solid line indicates the backgrounded dimension, while the dashed line shows the control dimensions. Error bars (offset for clarity) show 95% confidence intervals, with standard errors adjusted to reflect only between-subjects differences. The results for number of words used across features were similar and are not shown (see text for descriptives).

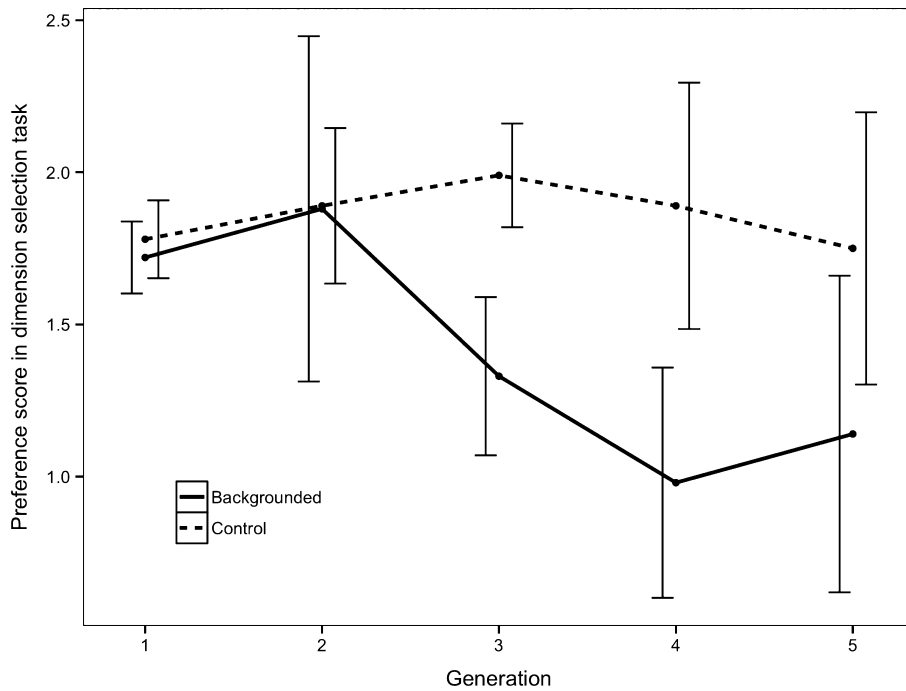


Fig. 5. Change in attention to different dimensions over generations, evaluated via the dimension selection task. The solid line indicates the backgrounded dimension, while the dashed line indicates the control dimensions. Error bars are 95% confidence intervals, with standard errors adjusted to reflect only between-subjects differences.

$SD = 0.13$), then gradually diverged over generations 2–5, with more words remaining on control dimensions than backgrounded dimensions. The greatest difference was in generation 4: $M_{\text{backgrounded}} = 1.94$, $SD = 0.40$; $M_{\text{control}} = 2.65$, $SD = 0.47$). Fixed effects of dimension salience, generation, and an interaction were included in the mixed-effects model. Analysis of this model showed that the main effect of dimension salience was significant, $\beta = 0.25$, $SE = 0.06$, $t(144) = 4.46$, $p < .001$. There was also a significant linear trend for the number of words to decrease over generations, $\beta = -0.92$, $SE = 0.11$, $t(144) = -8.29$, $p < .001$, and the effect of generation was also significantly different for backgrounded versus control dimensions, $\beta = 0.50$, $SE = 0.14$, $t(144) = 3.66$, $p < .001$.

Post-hoc tests (using the Bonferroni correction to establish a significance criterion of 0.008) found that the difference between backgrounded and control dimensions was marginally significant in generation 3, $t(7) = 3.54$, $p = .009$, and significant in generation 4, $t(7) = 4.03$, $p = .005$. The difference was not significant in any other generation ($t(7) < 1.71$, $p > .13$).

2. Within-dimension label dissimilarity (Fig. 4). Mean Levenshtein edit distance between words across backgrounded and control dimensions was similar at generation 1, then gradually diverged over generations 2–5. Words became more similar (i.e., edit distance was lower) on backgrounded dimensions than on control dimensions. The mixed-effects model incorporated fixed effects of dimension salience and generation, plus

an interaction. Analysis of this model showed that the main effect of dimension salience was significant overall, $\beta = 0.11$, $SE = 0.03$, $t(144) = 4.30$, $p < .001$. Additionally, word dissimilarity tended to decrease over generations, $\beta = -0.40$, $SE = 0.05$, $t(144) = -7.86$, $p < .001$, and the effect of generation was also significantly different for backgrounded versus control dimensions, $\beta = 0.19$, $SE = 0.06$, $t(144) = 2.97$, $p = .004$.

Post-hoc tests (using the Bonferroni correction to establish a significance criterion of .008) found that the difference between backgrounded and control dimensions was significant in generations 3, $t(7) = 3.92$, $p = .006$, and 4, $t(7) = 4.42$, $p = .003$. The difference was not significant in any other generation ($t(7) < 1.95$, $p > .09$).

3. Dimension selection task (Fig. 5). Mean selection preference score for backgrounded and control dimensions was similar at generations 1 and 2, then gradually diverged over generations 3–5, with higher preference scores on control dimensions than backgrounded dimensions. The mixed-effects model included fixed effects of dimension salience, generation, and an interaction. This model found significant main effects of dimension salience ($\beta = 0.45$, $SE = 0.12$, $t(3240) = 3.85$, $p < .001$) and generation ($\beta = -0.65$, $SE = 0.21$, $t(3240) = -3.14$, $p = .002$), and a significant interaction between the two ($\beta = 0.63$, $SE = 0.26$, $t(3240) = 2.42$, $p = .016$).

Post-hoc tests found a significant difference between backgrounded and control dimensions only in generations 3, $t(7) = 3.92$, $p = .006$, and 4, $t(7) = 3.70$, $p = .008$ (significance criterion using Bonferroni correction = .01). The difference was not significant at any other generation, $t(7) < 1.09$, $p > .11$.

4. Discussion

As predicted, patterns of underspecification that reflected the salience of dimensions in learning and production contexts arose gradually over generations of cultural transmission. Starting from input languages that specified equally across all dimensions, the languages lost distinctions earlier and faster on the dimension that was consistently backgrounded during learning and use. Fig. 6 shows a generation 5 language that underspecified more consistently on the backgrounded dimension (here, motion) than on the control dimensions. This was typical of the final languages in the experiment.

The gradualness of the effect is a product of individual-level learning processes amplified by cultural transmission. The first participant in each chain learns a language that sends a strong signal that all distinctions on all dimensions are important (since each image is labeled by a unique word). The performance of these participants on the dimension selection task shows that they have absorbed this expectation: They select all dimensions equivalently, showing that they consider them equally important to word meaning. However, this 27-word language is not learnable within the constraints of the training regime. Therefore, when these participants have to reproduce the language in the test phase, they are frequently faced with situations where they do not recall the word for the target referent. In this situation, a sensible strategy is to reuse a word they remember to

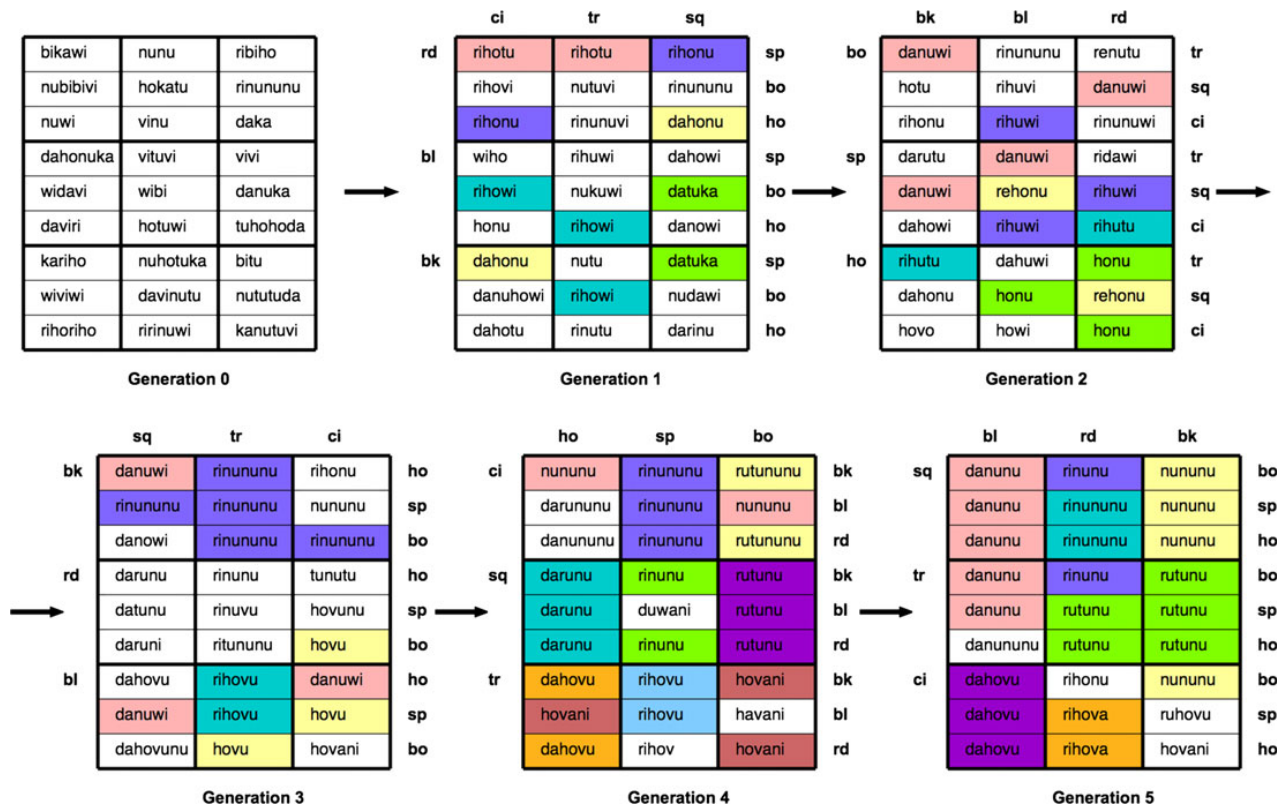


Fig. 6. The emergence of underspecification in chain 7. Each grid shows one participant's language, arranged so the backgrounded dimension always runs down the right-hand side. Abbreviations as in legend to Fig. 2. Words used for more than one referent, that is, underspecified words, are filled with the same color. The thick gridlines indicate regions that would be filled with the same color if the language underspecified on the backgrounded dimension. The figure shows underspecification arising more consistently on the backgrounded dimension than on either of the control dimensions, although it also extends partially to the control dimensions (see, e.g., the "overspill" of pink and green regions in generation 5).

be associated with at least one of the features of the referent. The question is, which feature(s) will they choose?

Globally, the initial language treats all dimensions as equally important. However, when participants are actually learning the meaning of each word, attending to the backgrounded dimension will never improve their success in the discrimination game. This systematic manipulation means that the learner will tend to associate words more reliably with their referents' features on the more salient control dimensions than on the less salient backgrounded dimension. The analogous systematic structure of the production task, where the participant is cued to produce a word that will successfully discriminate the target from the distractor, also influences him or her to use a word which he or she associates with a feature on the salient dimension(s), rather than on the backgrounded dimension.

Therefore, participants will tend to reuse words for multiple referents that differ on the backgrounded dimension. The participant's task is still to converge on the language they are trained on, so errors in this direction will tend to be small and not necessarily systematic. However, as these errors build up over generations, they change the language

and hence introduce a new source of evidence for the unimportance of one dimension: the patterns of word use in the language itself. Once a learner observes that a word can generalize over features on a dimension, this encourages the learner to reuse it for other cases if his or her memory fails (see Fig. 6 for a generation-by-generation view of how underspecification on the backgrounded dimension spreads as a chain progresses).

The majority of the output languages in generation 5 were underspecified across more than just the backgrounded dimension. The generation 5 language in Fig. 6 underspecifies not only across the motion dimension (the backgrounded dimension for this participant) but also partially across the shape dimension—for example, a blue square and a blue triangle are both called “danunu.” This shows that the undirected underspecification that arose in Kirby et al. (2008) also occurred in this experiment, in addition to the underspecification cued by the experimental manipulation. The learning- and testing-based cues in this experiment, while sufficient to direct underspecification preferentially toward backgrounded dimensions, are not sufficient to prevent it eventually arising on control dimensions. This expected outcome of iterated learning leads to the lack of a significant difference between backgrounded and control dimensions in generation 5, as detailed in the Results. In real language use, other pressures presumably prevent undirected underspecification from happening, for example a pressure for unambiguous communication. One avenue for future work is to explore whether, with the introduction of a pressure for unambiguous communication (following Smith, Tamariz, & Kirby, 2013), underspecification would still emerge on the backgrounded dimension, while distinctions on control dimensions would be preserved.

5. Conclusion

We set out to investigate how patterns of underspecification that help learners generalize words appropriately could arise in language. The results show that attentional learning effects amplified over cultural transmission lead to a lexicon that underspecifies preferentially across dimensions that are habitually less salient during learning and use. Thinking of the language in the experiment as analogous to a particular region of the lexicon, for example object or substance nouns, illuminates a possible mechanism for the origin of the strong tendencies to specify on particular dimensions we see in these regions. Over a whole language, specification will range over various dimensions depending on the function of individual words, as well as the characteristic situations in which they are learned and used. For example, the relational nature of gradable adjectives such as “big” means that the contexts in which they are learned and used will tend to highlight dimensions of relations between objects as well as intrinsic dimensions (Clark & Amaral, 2010; Gentner & Kurtz, 2005; Sandhofer & Smith, 2001). More broadly, a language can be seen as a dynamic system where the meanings of individual words adapt to, as well as themselves contributing to, the salience of particular dimensions in contexts of learning and use. This result uncovers a mechanism for how words can come to specify in adaptive ways: as a cumulative product of the incremental changes made by individual learners attending to contextual cues in learning and use.

Acknowledgments

CS is funded by an AHRC studentship.

Notes

1. Some of these factors concern the intrinsic salience of particular object features, rather than (as modeled in this experiment) task-defined salience in situations of learning and use. Intrinsic salience could also have a strong effect in directing underspecification, via the same mechanisms of cultural evolution modeled here.
2. To ensure that the payment of the last two chains of participants did not affect the results, Chain (i.e., which of the eight chains of five learners a participant belonged to) was modeled as a fixed effect in initial analyses to check if this improved the fit of the models. In all cases, the models including Chain as a fixed effect either did not improve overall fit or showed that no particular chain(s) had a significant effect on the results. In the final analyses below, Chain is modeled as a random effect.
3. The random effects to include in these models were assessed by means of likelihood ratio tests. All models incorporated a random intercept for Chain and a random slope for Participant.
4. $2 \times (1 - \text{pt}(\text{abs}(t), Y - Z))$, where Y is the number of observations, and Z is the number of fixed effect parameters. The `pt` command on R accesses the probability distribution for t . $Y - Z$ calculates the degrees of freedom, and multiplying by 2 obtains the p -value for a two-tailed test. Since this can be anticonservative at small sample sizes, we also used the heuristic of only accepting t values larger than 2 as significant (Baayen, 2008).

References

- Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics*. Cambridge, UK: Cambridge University Press.
- Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.
- Booth, A. E., & Waxman, S. (2002). Object names and object functions serve as cues to categories for infants. *Developmental Psychology*, 38(6), 948–957.
- Caldwell, C. A., & Smith, K. (2012). Cultural evolution and perpetuation of arbitrary communicative conventions in experimental microsocieties. *PLoS ONE*, 7(8), e43807.
- Clark, E. V. (1993). *The lexicon in acquisition*. Cambridge: Cambridge University Press.
- Clark, E. V. (1997). Conceptual perspective and lexical choice in acquisition. *Cognition*, 64(1), 1–37.
- Clark, E. V., & Amaral, P. M. (2010). Children build on pragmatic information in language acquisition. *Language and Linguistics Compass*, 4(7), 445–457.
- Cornish, H. (2011). *Language adapts: Exploring the cultural dynamics of iterated learning* (Unpublished doctoral dissertation). University of Edinburgh, Edinburgh, UK.
- Fay, N., Garrod, S., Roberts, L., & Swoboda, N. (2010). The interactive evolution of human communication systems. *Cognitive Science*, 34(3), 351–386.
- Geeraerts, D. (2009). *Theories of lexical semantics*. Oxford, England: Oxford University Press.

- Gentner, D., & Kurtz, K. J. (2005). Relational categories. In W. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. Wolff (Eds.), *Categorization inside and outside the laboratory: Essays in honor of Douglas L. Medin* (pp. 151–175). Washington, DC: American Psychological Association.
- Keil, F. C. (1994). Explanation, association, and the acquisition of word meaning. *Lingua*, *92*, 169–196.
- Kelemen, D., & Bloom, P. (1994). Domain-specific knowledge in simple categorization tasks. *Psychonomic Bulletin & Review*, *1*(3), 390–395.
- Kemler Nelson, D. (1995). Principle-based inferences in young children's categorization: Revisiting the impact of function on the naming of artifacts. *Cognitive Development*, *10*(3), 347–380.
- Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(31), 10681–10686.
- Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44.
- Lin, E. L., & Murphy, G. L. (1997). Effects of background knowledge on object categorization and part detection. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(4), 1153–1169.
- Markman, E. M. (1994). Constraints on word meaning in early language acquisition. *Lingua*, *92*, 199–227.
- Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207–238.
- Perfors, A., & Navarro, D. (2011). Language evolution is shaped by the structure of the world: An iterated learning analysis. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd annual conference of the Cognitive Science Society* (pp. 477–482). Austin, TX: Cognitive Science Society.
- Real, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, *111*(3), 317–328.
- Regier, T. (2005). The emergence of words: Attentional learning in form and meaning. *Cognitive Science*, *29*, 819–865.
- Sandhofer, C. M., & Smith, L. B. (2001). Why children learn color and size words so differently: Evidence from adults' learning of artificial terms. *Journal of Experimental Psychology: General*, *130*(4), 600–617.
- Smith, L. B., Colunga, E., & Yoshida, H. (2010). Knowledge as process: Contextually-cued attention and early word learning. *Cognitive Science*, *34*, 1287–1314.
- 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), 13–19.
- Smith, L. B., & Samuelson, L. (2006). An attentional learning account of the shape bias: Reply to Cimpian and Markman (2005) and Booth, Waxman, and Huang (2005). *Developmental Psychology*, *42*(6), 1339–1343.
- 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*(3), 444–449.
- Steels, L. (2003). Evolving grounded communication for robots. *Trends in Cognitive Sciences*, *7*(7), 308–312.
- Theisen-White, C., Kirby, S., & Oberlander, J. (2011). Integrating the horizontal and vertical cultural transmission of novel communication systems. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 956–961). Austin, TX: Cognitive Science Society.
- Tomasello, M. (2000). The social-pragmatic theory of word learning. *Pragmatics*, *10*(4), 401–413.
- Voiklis, J., & Corter, J. E. (2012). Conventional wisdom: Negotiating conventions of reference enhances category learning. *Cognitive Science*, *36*(4), 607–634.
- Waxman, S. R., & Kosowski, T. D. (1990). Nouns mark category relations: Toddlers' and preschoolers' word-learning biases. *Child Development*, *61*, 1461–1473.
- Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245–272.

Communication Leads to the Emergence of Sub-optimal Category Structures

Catriona Silvey (C.A.Silvey@sms.ed.ac.uk), Simon Kirby, Kenny Smith

Language Evolution and Computation Research Unit, School of Philosophy, Psychology and Language Sciences,
University of Edinburgh, Dugald Stewart Building,
3 Charles Street, Edinburgh, EH8 9AD, UK

Abstract

Words divide the world into labeled categories. Languages vary in the categories they label, sometimes to the point of making cross-cutting divisions of the same space. Previous work suggests two opposing hypotheses about how communication contributes to category emergence: 1) these spaces lack an objective shared similarity structure, and communication dynamically creates one of a number of optimally shareable category structures; 2) the category structures resulting from communication are not necessarily optimal, but diverge from a shared similarity space in language-specific ways. We had participants categorize images drawn from a continuous space in two conditions: a) non-communicative, by similarity, b) communicative, dynamically creating categories when playing a partnered communication game. The memory demands of communication lead to reliance on salient images and early conventions, resulting in non-optimal category structures compared to non-communicative participants. This supports the hypothesis that communication leads to categories that diverge non-optimally from a shared similarity space.

Keywords: communication; category structure; category emergence; language evolution

Introduction

Words divide the world into labeled categories. Languages vary in the categories they label, with some languages making coarser, finer, or even cross-cutting distinctions relative to how other languages carve up the same space (Bowerman & Choi, 2001; Malt, Sloman, & Gennari, 2003). Work is ongoing to quantify and classify this variation (Majid, Jordan, & Dunn, in progress). The mechanism by which a set of labeled categories emerges in a given language is however unclear. One hypothesis is that at least for some domains (e.g. spatial relations, containers), there is no one perceptually obvious way to divide the space into categories: there are several potential ways an individual observer could draw category boundaries (Bowerman, 2000). Some researchers have built on this idea to suggest that the process of communication itself structures a previously unstructured space, making categories that are optimally shareable between communicators (Freyd, 1983; Markman & Makin, 1998; Steels & Belpaeme, 2005; Voiklis & Corter, 2012). However, cross-linguistic work by Barbara Malt and colleagues on similarity perception versus labeling shows that, while the labeled categories of different languages do indeed diverge from each other, speakers of different languages still perceive the similarities between the objects in comparable ways (Malt, Sloman, Gennari, Shi, & Wang, 1999). This suggests that the categorization systems of different languages can in fact superimpose a range of divergent structures on a space that has a shared underlying similarity structure. These two accounts suggest radically different roles for communication in the emergence of categories.

The current experiment contributes to this debate by investigating how humans categorize a set of images designed to have unclear category boundaries. The participants categorize the images in one of two conditions: a non-communicative condition, where solo participants divide the images into categories according to similarity, and a communicative condition, where pairs of participants play a communication game with the images. The results shed light on the effect of communication on category structure, suggesting that the categories created by communication can and do diverge from a relatively shared similarity space, even in a stimulus set designed to have ambiguous boundaries.

Method

Participants were assigned to two conditions. In the non-communicative condition, participants divided a continuous space of images into labeled categories on the basis of similarity. In the communicative condition, pairs of participants played a communication game using the same continuous space of images. Participants in this condition produced labeled categories via the words they used to communicate each target image in the last two rounds of the experiment. The category systems the participants produced in the two conditions were then compared.

Stimuli

The set of images used in the experiment is shown in Figure 1. The four corner images were generated using PsychoPy software (Peirce, 2007). For each image, a random number generator assigned x and y positions for the five vertices, and the resulting shape was drawn. Morphs between these images were then generated by shifting the vertices towards each of the corners, according to a weight defined by inverse Euclidean distance (Matthews, 2009), to create a total set of 25 images. The 'objective' Euclidean distance between the images in the space may of course not correspond to perceptual similarity (see, e.g., Smith & Heise, 1992); however, in pilot experiments, participants showed variation in where they drew the category boundaries, making these stimuli suitable for the current study.

Labels

To control for any effects on participants' categorizations arising purely from the use of labels (Lupyan, Rakison, & McClelland, 2007), words to label the categories were provided in both the non-communicative and communicative conditions. Lists of 25 CVCV nonsense words were generated by combining consonants and vowels randomly selected

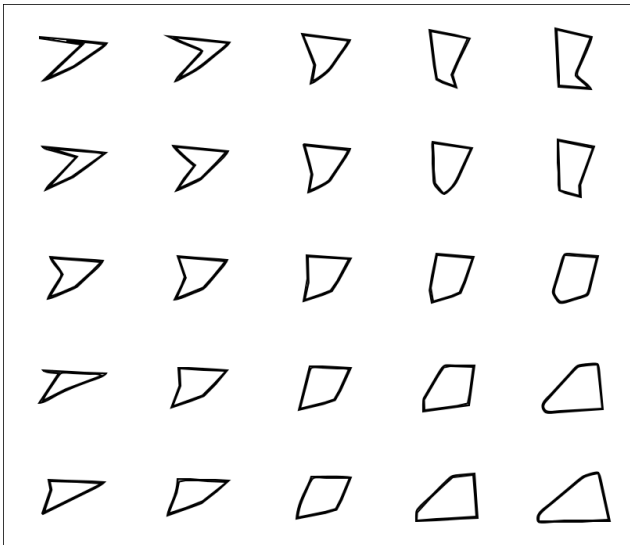


Figure 1: The stimuli used in the study (lines thickened for clarity).

from the whole alphabet (e.g., *zipi*, *gisa*, *wada*). Since we expected that participants would use known crossmodal associations between attributes of words and attributes of the images in assigning category labels (e.g. voiceless stops and spikiness, Nielsen & Rendall, 2011), we assigned the same wordlist to a yoked triple of two non-communicative participants and one communicative pair, so that in the analyses, any peculiar effects of a particular wordlist would apply equally across the conditions.

Participants

Participants were 42 students at the University of Edinburgh (30 female, median age 23). 20 took part in the non-communicative condition. The non-communicative experiment took 15 minutes. Participants were paid £2. 22 participants (randomly assigned into 11 pairs) took part in the communicative condition. The communicative experiment took an hour. Participants were paid £7, and each member of the pair with the highest communication score was awarded a £10 Amazon voucher. One pair failed to complete the experiment within an hour and so was excluded from analyses.

Procedure

Non-Communicative Condition Participants were presented with a randomized onscreen array of all 25 images and a set of words to label categories. To avoid cueing the participants to produce a particular number of categories, only one word was initially shown on screen: participants could reveal new words at any time, and were told that a) they could use as few or as many words as they wanted, and b) they did not have to use all the words they had revealed. Participants could reveal a new word at any stage, up to 25 words. They were instructed to label similar images with the same word and different images with different words.

Communicative Condition Participants communicated via computer terminals in separate cubicles. In a communication trial, one participant was assigned as the sender and one as the receiver. The sender was presented with a randomized onscreen array of all 25 images, one of which was selected with a red box to indicate it was the target. The sender was also presented with one initial word. The sender could reveal a new word at any stage, up to 25 words. Any words they had revealed on a previous trial remained visible on their screen for all subsequent trials. The participant was instructed to choose a word that would help the receiver pick out the target from the array of images.

Once the sender had picked a word, the receiver was presented with a randomized onscreen array of all 25 images and the word the sender had chosen. The receiver was instructed to select the image the sender had wanted to communicate.

Once the receiver selected an image, both participants were presented with a feedback screen. The feedback screen showed the word the sender had used, the target image, the image the receiver had selected, the score for the trial, and the running score for the whole experiment. The score for each trial was calculated on the basis of the inverse Euclidean distance between the target and the image the receiver selected, from a minimum of 1 up to a maximum of 15 (for correctly picking the target).

After each communication trial the sender and the receiver swapped roles. The experiment consisted of 100 communication trials divided into 4 rounds. Each round featured the 25 images as targets in a randomized order. The randomized lists were balanced such that each participant was the sender for every target image once in the first half of the experiment, and once in the second half.

The first two rounds of the experiment were not incorporated into the categorization analysis, as it was expected that at this stage a system would still be emerging. Participants' categories were therefore taken from the last two rounds of the experiment. Success scores were taken from the whole experiment.

Dependent Variables

Number of Categories The number of categories each participant produced was recorded.

Variation in Category Size To achieve a measure of variation in category size that took the number of categories into account (since more categories would generally contain fewer images each), the number of images in each category was divided by the expected number of images in each category, if images were distributed equally. For example, if a participant had 5 categories, an equal distribution would be to place 5 images in each category: if one of their categories in fact had 10 images, this would produce a value for that category of $10/5 = 2$. The range of these values was then taken as a measure of variation in category size adjusted for the number of categories (with a minimum value of 0 in the case of perfectly balanced categories).

Category Alignment Two measures were taken to compare participants' categories and quantify their alignment. The first, the Rand index (Rand, 1971), consists of a pairwise comparison of whether participants tended to place images in the same category or different categories. The calculation produces a value bounded from 0 to 1, where 1 is perfect alignment. The second, V-Measure (Rosenberg & Hirschberg, 2007), is based on variation of information between the groupings, normalized to compensate for differences in number of categories. This measure also ranges from 0 to 1 where 1 is perfect alignment. Two further measures, the Variation of Information measure on which V-Measure is based (Meilä, 2003) and an adjusted version of Cramer's phi (Wills & McLaren, 1998) were considered, but were found to produce incongruent results when applied to groupings with divergent numbers of categories. Since the variable of interest was participants' categories rather than the words they used, the alignment measures were taken irrespective of whether participants used the same words: i.e., if two participants put the same set of images in a labeled category but used different labels, they would count as fully aligned for this category.

Hypotheses

For the non-communicative participants, there is no particular incentive to divide the images into more or fewer categories (beyond the minimal assumption that, in being asked to sort the images, the participants are unlikely to place them all in one category). This condition therefore functions as a baseline for assessing the variability of the participants' categorization of the images without communication. The expectation is that with no strong motivation to behave in any particular way, participants' categorization performance will vary.

For the communicative participants, the pressures on their emergent categorization systems are more complex. The only way to attain a perfect communication score with this stimulus space and scoring system is to have a unique label for each image, i.e. 25 words in total, with 25 corresponding categories containing one image each. However, participants' memory constraints will likely prevent this from happening in the experiment. More generally, then, for a given number of words, the optimal strategy is to apply each word to an equal number of images in the space, in a contiguous region (Gärdenfors, 2000). Participants who converge on a system like this would maximize their possible score across all rounds of communication. Figure 2A shows an example of this kind of optimal system. When the sender uses a word corresponding to one of the categories, the receiver can adopt the strategy of picking a central member of the category, thus ensuring their response is a maximum of 1.4 Euclidean distance units (or one diagonal step) from the target. Figure 2B shows, by contrast, a non-optimal system with the same number of categories. This system is non-optimal for two reasons. 1) The number of images in each category is less balanced (one category contains only two images, while another contains ten). This means that when the sender uses the word for the bigger category, the probability of the receiver selecting

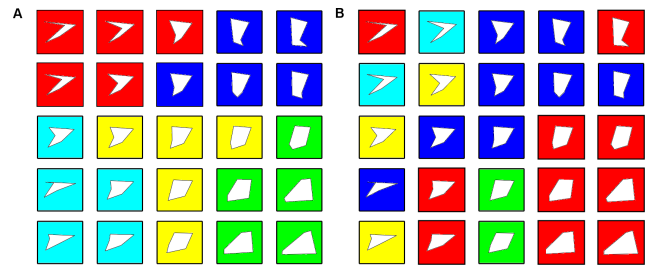


Figure 2: A) An example of a category system optimally structured for communicative success. B) A non-optimal system with the same number of categories.

an image close to the target is lower. 2) The images belonging to some categories are spread across different regions of the space and do not form contiguous regions. This raises the probability of a receiver selecting an image some distance away from the target, even if she shares this set of categories with the sender. It is worth noting that the spaces we categorize in the real world may not have this kind of smooth continuous structure, and so the regular contiguous regions of Figure 2A may be more difficult to achieve. However, in the context of this experiment, if communication does give rise to optimally structured categories, this is the kind of system we would expect to see emerging.

Results

A linear trend ANOVA found that communicative success increased over the 4 rounds of the experiment, $F(1, 9) = 18.66$, $p = .002$ (Figure 3). Participants' overall success was significantly above chance, $t(9) = 4.21$, $p = .002$.

Participants in the communicative condition used significantly more labeled categories ($M = 9.95$, $SD = 3.98$) than participants in the non-communicative condition ($M = 6$, $SD = 1.37$), Mann-Whitney $U = 60$, $z = -3.54$, $p < .001$. Communicative participants also showed significantly more variance in how many labeled categories they used, Levene's test $(1, 36) = 16.47$, $p < .001$. Pairs who communicated together, however, showed no significant difference in the number of categories they used, $t(18) = -0.38$, $p = .7$, showing that this effect came from differences between, rather than within, communicative pairs. Thus, even though the non-communicative participants had less motivation to converge on a particular number of labeled categories, they were more consistent in the number they produced than the communicative participants.

Participants in the communicative condition also varied significantly more in the size of their categories, when number of categories was taken into account (category size variation as described in Methods $M = 1.54$, $SD = 0.35$, compared to non-communicative participants, $M = 1.17$, $SD = 0.4$). That is, images were more unevenly distributed across categories in the communicative condition, $t(38) = 3.13$, $p < .005$. Surprisingly for the communication-as-alignment hypothesis, communicative pairs' groupings did not align sig-

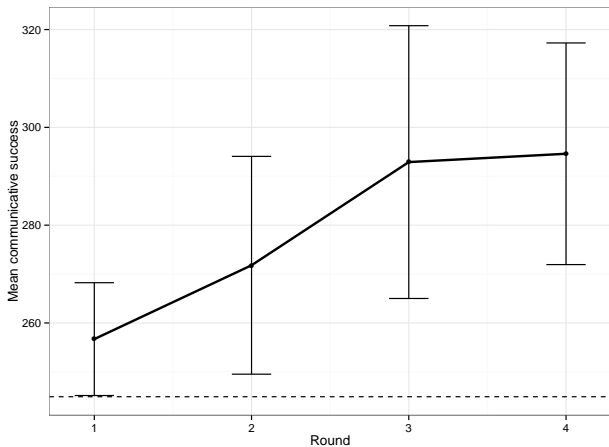


Figure 3: Average communicative success over rounds in the experiment. Dotted line shows chance. Error bars show 95% confidence intervals.

nificantly more than non-communicative participants' (by-language analysis: paired-samples t -test $0.47 < t(9) < 0.63$, $p > .5$, by-subjects analysis: independent t -test $-0.42 < t(18) < 0.63$, $p > .4$). Neither did communicative success correlate significantly with either of the alignment measures, $r < .51$, $p > .14$.

To test the hypothesis that communicative participants within a pair were more aligned than communicative participants who were not paired with each other, an analysis was run comparing the alignment scores for the true pairs with alignment scores for shuffled pairs (participant 2 paired with participant 3, etc.). A similar analysis was run for the non-communicative pairs, comparing alignment of those who shared the same wordlist with those who had different wordlists. Non-communicative participants displayed equivalent levels of alignment whether or not they used the same wordlist, $t(9) < 0.8$, $p > .58$. For communicative participants, one of the alignment measures (Rand index) tended towards being significantly higher for participants who communicated in a pair than participants who did not, $t(9) = 1.88$, $p = .093$, suggesting that communicative participants were marginally more aligned within-pair than between-pair in terms of which pairs of images they categorized together. For the second alignment measure, V-Measure, no significant difference was found, $t(9) = 1.22$, $p > .25$.

Discussion

The results are somewhat surprising for the hypothesis that communication creates optimal structure in previously variably structured spaces. Communicative participants produced categorizations that were generally non-optimal for maximizing communicative success, as defined in Hypotheses above. This is not merely a property of how humans perceive this particular space, as shown by the contrast with the non-communicative condition, where participants' categories

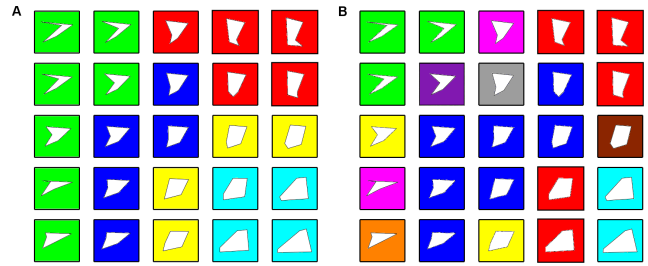


Figure 4: A) A typical non-communicative participant's categories. B) A typical communicative participant's categories.

were generally more balanced in size, carving up the space in a way that would actually be more optimal by this definition. Figure 4 shows a typical example of A) a non-communicative participant's categories and B) a communicative participant's categories. It is notable that several categories in B are also non-optimal in that they cover non-contiguous regions of the space (e.g. red and yellow categories). The heatmaps in Figures 5 and 6 show more generally how communicative participants' categories were more dispersed (Figure 6) compared to non-communicative participants, who tend to clump more around certain pairings or groups to form their categories (darker regions in Figure 5).

Why did communicative participants divide up the space so differently from non-communicative participants? As mentioned in Hypotheses above, the communicative task exerts a considerable memory demand on participants: although they are presented with the full image space on each trial, they still have to remember which word applies to which image or group of images over the course of the experiment. This exerts a pressure to create a system that is optimized not just for communicative success, but also for learnability.

Aids to learnability in this experiment might include particularly salient words, images, and pairings between them, or felicitous early successes that lead to the forming of conventions. These conventions, once established, may then prove too valuable to shift in favor of more optimally structured categories. Both of these aids to learnability (salient images/words and early successes) are mentioned by participants in the post-experiment questionnaire. Typically, when asked to draw the images they remember, participants could draw from memory two to five salient images and their associated words, but were unclear on other regions of the space. Thus the memory demands of the task, and the fact that participants have to establish a system from scratch, make the salience of individual images and early established conventions important factors determining the shape of each participant's final categorization system.

The possibility that different images in the set had differing salience is also supported by the success heatmaps in Figure 7. The heatmap in Figure 7A shows which target images led to higher success scores for participants. The pattern here is at odds with Figure 7B, which shows the relative expected

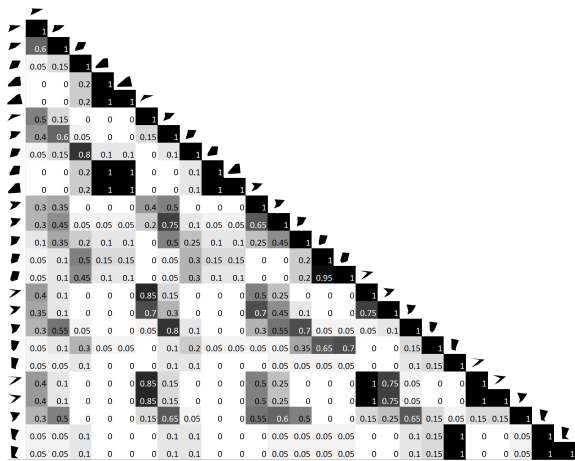


Figure 5: Heatmap visualizing how often non-communicative participants placed pairs of images in the same category. Darker areas indicate pairs more often categorized together.

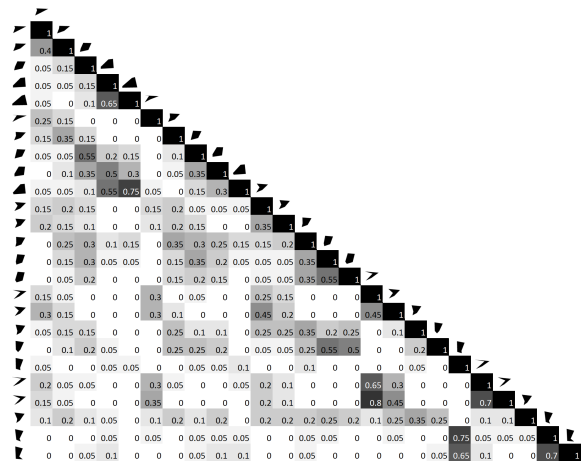


Figure 6: Heatmap visualizing how often communicative participants placed pairs of images in the same category. Darker areas indicate pairs more often categorized together.

chance levels of success for each image: images in the middle have more low-ED neighbors, so the probability of a higher score goes up when they are the target. The fact that panels A and B differ shows that participants' success with particular images is boosted by some other factor.

Panel C shows a heatmap of this boost – darker images are those whose overall communicative success rate is highest compared to what the chance-based map in panel B would predict. The likely explanation is that these images have higher salience for participants, making them act as Schelling points between sender and receiver. The striking finding that communicative success is not correlated with overall alignment could therefore be explained by participants consolidating success on a few images, leaving other areas of the space more sparsely covered.

While Figure 7 suggests that the salience of particular images may be shared across all pairs, early conventions are

more likely to vary between pairs due to the randomized presentation of targets. This could explain the tendency towards higher pairwise (Rand index) alignment within pairs than between pairs, as reported in the Results. Despite the low levels of alignment overall, communicative pairs' language-specific early conventions may bring them more into agreement on how they categorize specific small groups of images.

As mentioned above, the pressures on the participants in the two conditions were substantially different: participants in the non-communicative condition interacted with the stimuli more briefly and without memory constraints, as well as lacking the pressure to create more categories imposed by the communicative task. Future work could investigate how participants divide up the space non-communicatively under the same time and memory constraints as the communicative participants, thus disentangling the effects of these constraints from the effects of communication. The non-communicative condition in this study still serves as a useful

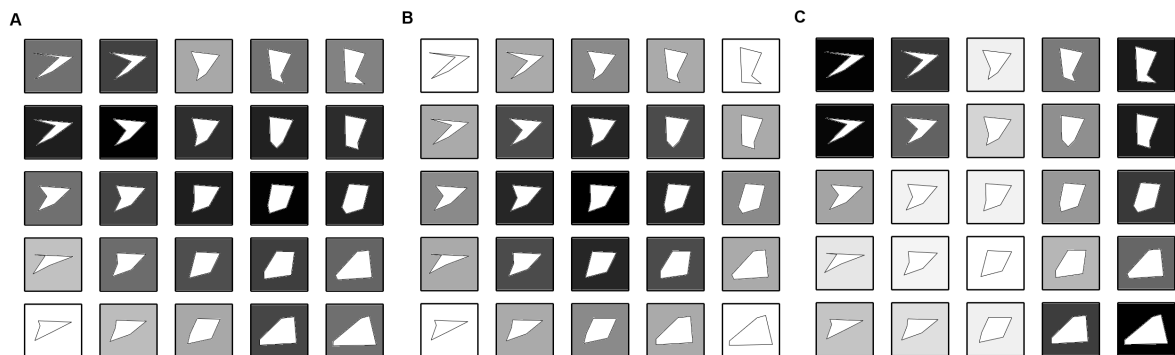


Figure 7: Heatmaps showing which target images produced higher per-round success scores. Darker images produced higher scores. A) Map of overall success per image in the experiment. B) Map of expected chance success rates per image. C) Difference between maps A and B. Darker images are those whose success rates are boosted highest beyond expected.

baseline, however, for participants' perceptually based divisions of the space.

The outcome of this study – that communication does not necessarily optimize category structures, but can create uneven and suboptimal structures compared to non-communicators' division of the same space – is reflected in our experience of real language, where words vary widely in whether they specify small regions of semantic space or broad undifferentiated regions. The existence of the latter kind of word does not necessarily mean the users of the language do not perceive the differences between sub-parts of the region it covers: only that, for reasons of salience, or constraints imposed by the history and development of conventions in the language, these internal differences lack category labels. An important additional pressure in real language, not modeled in this study, is that different regions of semantic space may also have different functional importance, motivating coarser or finer-grained distinctions in different regions. However, these results show that even in the absence of functional reasons for uneven division of a space, communication can lead to the establishment of categories that may not align with non-communicative similarity perception.

Conclusion

Communication is not a simple process of mapping words onto pre-shared perceptual categories. Even if communicating partners agree on the underlying structure of the space they are talking about, the categories that emerge from communication can diverge in surprising ways, both from the underlying similarity space and from the category structure that would be most optimal for communicative success. Constraints on learning, salience effects, and the impact of early conventions on a language's development all contribute to shaping an emergent system of labeled categories.

Acknowledgments

CS is supported by an AHRC PhD studentship. Thanks to Christos Christodoulopoulos for help with alignment measures and with networking for the communication experiment, and to Mark Atkinson and Andrea Ravignani for help with piloting.

References

- Bowerman, M. (2000). Where do children's word meanings come from? Rethinking the role of cognition in early semantic development. In L. Nucci, G. Saxe, & E. Turiel (Eds.), *Culture, thought and development* (pp. 199–230). Mahwah, NJ: Lawrence Erlbaum.
- Bowerman, M., & Choi, S. (2001). Shaping meanings for language: Universal and language-specific in the acquisition of spatial semantic categories. In M. Bowerman & S. C. Levinson (Eds.), *Language acquisition and conceptual development* (pp. 475–511). Cambridge: Cambridge University Press.
- Freyd, J. (1983). Shareability: The social psychology of epistemology. *Cognitive Science*, 7(3), 191–210.
- Gärdenfors, P. (2000). *Conceptual spaces: The geometry of thought*. Cambridge, MA: MIT Press.
- Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not just for talking: Redundant labels facilitate learning of novel categories. *Psychological Science*, 18(12), 1077–83.
- Majid, A., Jordan, F., & Dunn, M. (in progress). *Evolution of semantic systems*. <http://www.mpi.nl/departments/other-research/research-consortia/eoss>. (Online; accessed 22 April 2013)
- Malt, B. C., Sloman, S. A., Gennari, S., Shi, M., & Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language*, 40(2), 230–262.
- Malt, B. C., Sloman, S. A., & Gennari, S. P. (2003). Universality and language specificity in object naming. *Journal of Memory and Language*, 49(1), 20–42.
- Markman, A. B., & Makin, V. S. (1998). Referential communication and category acquisition. *Journal of Experimental Psychology: General*, 127(4), 331–54.
- Matthews, C. (2009). *The emergence of categorization: Language transmission in an iterated learning model using a continuous meaning space*. Unpublished master's thesis, University of Edinburgh.
- Meilä, M. (2003). Comparing clusterings by the variation of information. In B. Schölkopf & M. K. Warmuth (Eds.), *Learning theory and kernel machines* (pp. 173–187). Berlin: Springer-Verlag.
- Nielsen, A., & Rendall, D. (2011). The sound of round: Evaluating the sound-symbolic role of consonants in the classic Takete-Maluma phenomenon. *Canadian Journal of Experimental Psychology*, 65(2), 115–24.
- Peirce, J. W. (2007). PsychoPy—psychophysics software in Python. *Journal of Neuroscience Methods*, 162(1), 8–13.
- Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336), 846–850.
- Rosenberg, A., & Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* (pp. 410–420).
- Smith, L. B., & Heise, D. (1992). Perceptual similarity and conceptual structure. In B. Burns (Ed.), *Percepts, concepts and categories* (pp. 233–272). Amsterdam: Elsevier B.V.
- Steels, L., & Belpaeme, T. (2005). Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and Brain Sciences*, 28, 469–529.
- Voiklis, J., & Corter, J. E. (2012). Conventional wisdom: Negotiating conventions of reference enhances category learning. *Cognitive Science*, 36(4), 607–634.
- Wills, A. J., & McLaren, I. P. L. (1998). Perceptual learning and free classification. *The Quarterly Journal of Experimental Psychology*, 51B(3), 235–270.

Bibliography

- Abelson, R. P., & Sermat, V. (1962). Multidimensional scaling of facial expressions. *Journal of Experimental Psychology*, *63*(6), 546–54.
- Akmajian, A., Demers, R. A., Farmer, A. K., & Harnish, R. M. (2001). *Linguistics*. Cambridge, MA: MIT Press.
- Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–64.
- Altmann, G. T. M., & Mirković, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science*, *33*(4), 583–609.
- Arbib, M. A. (2005). From monkey-like action recognition to human language: an evolutionary framework for neurolinguistics. *Behavioral and Brain Sciences*, *28*(2), 105–24; discussion 125–67.
- Arbib, M. A., Liebal, K., & Pika, S. (2008). Primate vocalization, gesture, and the evolution of human language. *Current Anthropology*, *49*(6), 1053–1076.
- Arnold, K., & Zuberbühler, K. (2006). Language evolution: semantic combinations in primate calls. *Nature*, *441*(7091), 303.
- Au, T. K. (1983). Chinese and English counterfactuals: the Sapir-Whorf hypothesis revisited. *Cognition*, *15*(1-3), 155–87.
- Aydede, M. (2010). The language of thought hypothesis. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. Fall 2010 ed.
- Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics*. Cambridge: Cambridge University Press.

- Bergman, T. J., Beehner, J. C., Cheney, D. L., & Seyfarth, R. M. (2003). Hierarchical classification by rank and kinship in baboons. *Science*, *302*(5648), 1234–6.
- Bickerton, D. (1990). *Language and species*. Chicago: Chicago University Press.
- Bicknell, K., Elman, J. L., Hare, M., McRae, K., & Kutas, M. (2010). Effects of event knowledge in processing verbal arguments. *Journal of Memory and Language*, *63*(4), 489–505.
- Bloom, A. H. (1981). *The linguistic shaping of thought: A study in the impact of language on thinking in China and the West*. Hillsdale, NJ: Erlbaum.
- Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.
- Bloom, P., & Keil, F. C. (2001). Thinking through language. *Mind & Language*, *16*(4), 351–367.
- Booth, A. E., & Waxman, S. (2002). Object names and object functions serve as cues to categories for infants. *Developmental Psychology*, *38*(6), 948–957.
- Bowerman, M. (2000). Where do children's word meanings come from? Rethinking the role of cognition in early semantic development. In L. Nucci, G. Saxe, & E. Turiel (Eds.) *Culture, thought and development*, chap. 9, (pp. 199–230). Mahwah, NJ: Lawrence Erlbaum.
- Bowerman, M., & Choi, S. (2001). Shaping meanings for language: universal and language-specific in the acquisition of spatial semantic categories. In M. Bowerman, & S. C. Levinson (Eds.) *Language acquisition and conceptual development*, (pp. 475–511). Cambridge: Cambridge University Press.
- Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1482–1493.
- Brighton, H. (2002). Compositional syntax from cultural transmission. *Artificial Life*, *8*(1), 25–54.
- Brock, J., & Nation, K. (2014). The hardest butter to button: immediate context effects in spoken word identification. *Quarterly Journal of Experimental Psychology*, *67*(1), 114–23.

- Burling, R. (2007). *The talking ape: How language evolved*. Oxford: Oxford University Press.
- Burton, M. L. (2003). Too many questions? The uses of incomplete cyclic designs for paired comparisons. *Field Methods*, *15*(2), 115–130.
- Cacciari, C., & Tabossi, P. (1988). The comprehension of idioms. *Journal of Memory and Language*, *27*, 668–683.
- Caramazza, A., & Grober, E. (1976). Polysemy and the structure of the subjective lexicon. In C. Rameh (Ed.) *Georgetown University roundtable on languages and linguistics. Semantics: Theory and application*, (pp. 181–206). Washington, DC: Georgetown University Press.
- Carpenter, M., Nagell, K., Tomasello, M., Butterworth, G., & Moore, C. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development*, *63*(4), 1–174.
- Carston, R. (2007). Linguistic communication and the semantics/pragmatics distinction. *Synthese*, *165*(3), 321–345.
- Cartmill, E. A., Roberts, S., Lyn, H., & Cornish, H. (2014). *The Evolution of Language: Proceedings of the 10th International Conference (EVO LANG10)*. Singapore: World Scientific.
- Casasanto, D., & Lupyan, G. (in press). All concepts are ad hoc concepts. In E. Margolis, & S. Laurence (Eds.) *Concepts: New directions*. Cambridge, MA: MIT Press.
- Chambers, C. C., Tanenhaus, M. K., Eberhard, K. M., Filip, H., & Carlson, G. N. (2002). Circumscribing referential domains during real-time language comprehension. *Journal of Memory and Language*, *47*(1), 30–49.
- Cheney, D. L., & Seyfarth, R. M. (1985). Vervet monkey alarm calls: Manipulation through shared information? *Behaviour*, *94*(1/2), pp. 150–166.
- Choi, S., & Bowerman, M. (1991). Learning to express motion events in English and Korean: the influence of language-specific lexicalization patterns. *Cognition*, *41*(1-3), 83–121.

- Choi, S., McDonough, L., Bowerman, M., & Mandler, J. M. (1999). Early sensitivity to language-specific spatial categories in English and Korean. *Cognitive Development, 14*, 241–268.
- Chouinard, M. M., & Clark, E. V. (2003). Adult reformulations of child errors as negative evidence. *Journal of Child Language, 30*(3), 637–669.
- Christy, T. C. (1983). *Uniformitarianism in linguistics*. Amsterdam: John Benjamins.
- Clark, E. V. (1973). Non-linguistic strategies and the acquisition of word meanings. *Cognition, 2*(2), 161–182.
- Clark, E. V. (1993). *The lexicon in acquisition*. Cambridge: Cambridge University Press.
- Clark, E. V. (1997). Conceptual perspective and lexical choice in acquisition. *Cognition, 64*(1), 1–37.
- Clark, E. V. (2003). *First language acquisition*. Cambridge: Cambridge University Press.
- Clark, E. V. (2004). How language acquisition builds on cognitive development. *Trends in Cognitive Sciences, 8*(10), 472–8.
- Clark, E. V. (2007). Conventionality and contrast in language and language acquisition. *New Directions for Child and Adolescent Development, 115*, 11–23.
- Clark, E. V., & Amaral, P. M. (2010). Children build on pragmatic information in language acquisition. *Language and Linguistics Compass, 4*(7), 445–457.
- Clark, H. H. (1996). *Using language*. Cambridge: Cambridge University Press.
- Cohen, L. J. (1986). How is conceptual innovation possible? *Erkenntnis, 25*(2), 221–238.
- Collier, K., Bickel, B., van Schaik, C. P., Manser, M. B., & Townsend, S. W. (2014). Language evolution: syntax before phonology? *Proceedings of the Royal Society B: Biological Sciences, 281*, 20140263.
- Cornish, H. (2011). *Language adapts: Exploring the cultural dynamics of iterated learning*. PhD, University of Edinburgh.

- Croft, W. (1998). Linguistic evidence and mental representations. *Cognitive Linguistics*, 9(2), 151–173.
- Croft, W., & Cruse, D. A. (2004). *Cognitive linguistics*. Cambridge: Cambridge University Press.
- Cromer, R. F. (1974). The development of language and cognition: The cognition hypothesis. In B. Foss (Ed.) *New perspectives in child development*, (pp. 184–252). Harmondsworth: Penguin.
- Cruse, D. A. (2011). *Meaning in language: An introduction to semantics and pragmatics*. Oxford: Oxford University Press, 3 ed.
- Cumming, G. (2012). *Understanding the new statistics: effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.
- Cumming, S. (2013). Names. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. Spring 2013 ed.
- Cuskley, C., & Kirby, S. (2013). Synaesthesia, cross-modality and language evolution. In J. Simner, & E. M. Hubbard (Eds.) *Oxford Handbook of Synaesthesia*, (pp. 869–907). Oxford: Oxford University Press.
- Dahan, D., & Tanenhaus, M. K. (2004). Continuous mapping from sound to meaning in spoken-language comprehension: immediate effects of verb-based thematic constraints. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 30(2), 498–513.
- David, H. A. (1963). The structure of cyclic paired-comparison designs. *Journal of the Australian Mathematical Society*, 3, 117–127.
- de Boer, B., Sandler, W., & Kirby, S. (2012). New perspectives on duality of patterning: Introduction to the special issue. *Language and Cognition*, 4(4), 251–260.
- Deacon, T. W. (1997). *The symbolic species: The coevolution of language and the brain*. New York: Norton.
- Deacon, T. W. (2003). Universal grammar and semiotic constraints. In M. H. Christiansen, & S. Kirby (Eds.) *Language Evolution*, chap. 7, (pp. 111–139). Oxford: Oxford University Press.

- Deutscher, G. (2005). *The unfolding of language*. London: Arrow.
- Dolscheid, S., Shayan, S., Majid, A., & Casasanto, D. (2013). The thickness of musical pitch: psychophysical evidence for linguistic relativity. *Psychological Science*, 24(5), 613–21.
- Dunbar, R. (1996). *Grooming, gossip and the evolution of language*. London: Faber & Faber.
- Ekman, G. (1954). Dimensions of color vision. *The Journal of Psychology*, 38, 467–474.
- Elbourne, P. (2011). *Meaning: A slim guide to semantics*. Oxford: Oxford University Press.
- Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science*, 33(4), 547–582.
- Estill, R. B., & Kemper, S. (1982). Interpreting idioms. *Journal of Psycholinguistic Research*, 11(6), 559–568.
- Evans, V. (2009). *How words mean: Lexical concepts, cognitive models, and meaning construction*. Oxford: Oxford University Press.
- Everitt, B. S., & Hothorn, T. (2010). *A handbook of statistical analyses using R*. Boca Raton, FL/London: Chapman & Hall.
- Fagot, J., Wasserman, E. A., & Young, M. E. (2001). Discriminating the relation between relations: The role of entropy in abstract conceptualization by baboons (*Papio papio*) and humans (*Homo sapiens*). *Journal of Experimental Psychology: Animal Behavior Processes*, 27(4), 316–328.
- Fillmore, C. J. (1970). The grammar of hitting and breaking. In R. Jacobs, & P. Rosenbaum (Eds.) *Readings in English transformational grammar*, (pp. 120–133). Waltham, MA: Ginn.
- Fitch, W. T. (2010). *The evolution of language*. Cambridge: Cambridge University Press.
- Fodor, J. A. (1981). The present status of the innateness controversy. In *Representations*, (pp. 257–333). Cambridge, MA: MIT Press.

- Fodor, J. A. (1998). *Concepts: Where cognitive science went wrong*. Oxford: Oxford University Press.
- Fodor, J. A., Garrett, M. F., Walker, E. C., & Parkes, C. H. (1980). Against definitions. *Cognition*, 8(3), 263–7.
- Folia, V., Udd, J., Forkstam, C., & Petersson, K. M. (2010). Artificial language learning in adults and children. *Language Learning*, 60(Suppl. 2), 188–220.
- Freyd, J. (1983). Shareability: The social psychology of epistemology. *Cognitive Science*, 7(3), 191–210.
- Galantucci, B. (2005). An experimental study of the emergence of human communication systems. *Cognitive Science*, 29(5), 737–67.
- Gärdenfors, P. (2000). *Conceptual spaces: the geometry of thought*. Cambridge, MA: MIT Press.
- Garrod, S., & Doherty, G. (1994). Conversation, co-ordination and convention: An empirical investigation of how groups establish linguistic conventions. *Cognition*, 53, 181–215.
- Garrod, S., Fay, N., Lee, J., Oberlander, J., & MacLeod, T. (2007). Foundations of representation: Where might graphical symbol systems come from? *Cognitive Science*, 31(6), 961–87.
- Garrod, S., & Pickering, M. J. (2009). Joint action, interactive alignment, and dialog. *Topics in Cognitive Science*, 1(2), 292–304.
- Geeraerts, D. (1993). Vagueness's puzzles, polysemy's vagaries. *Cognitive Linguistics*, 4(3), 223–272.
- Gennari, S., & Poeppel, D. (2003). Processing correlates of lexical semantic complexity. *Cognition*, 89(1), B27–B41.
- Gentner, D., & Kurtz, K. J. (2005). Relational categories. In W.-k. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. Wolff (Eds.) *Categorization inside and outside the laboratory: Essays in Honor of Douglas L. Medin*, chap. 9, (pp. 151–175). Washington, D.C.: American Psychological Association.

- Gibbs, R. W. (1985). On the process of understanding idioms. *Journal of Psycholinguistic Research*, 14(5), 465–472.
- Gibbs, R. W., Nayak, N. P., & Cutting, C. (1989). How to kick the bucket and not decompose : Analyzability and idiom processing. *Journal of Memory and Language*, 28, 576–593.
- Glucksberg, S., Kreuz, R. J., & Rho, S. H. (1986). Context can constrain lexical access: Implications for models of language comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 12(3), 323–335.
- Goldin-Meadow, S. (2007). Pointing sets the stage for learning language—and creating language. *Child Development*, 78(3), 741–5.
- Goldin-Meadow, S., Gelman, S. A., & Mylander, C. (2005). Expressing generic concepts with and without a language model. *Cognition*, 96(2), 109–26.
- Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology*, 123(2), 178–200.
- Gómez, J.-C. (2007). Pointing behaviors in apes and human infants: a balanced interpretation. *Child Development*, 78(3), 729–34.
- Grice, H. P. (1957). Meaning. *The Philosophical Review*, 66(3), 377–388.
- Grice, H. P. (1975). Logic and conversation. In P. Cole, & J. L. Morgan (Eds.) *Syntax and Semantics*, (pp. 41–58). New York: Academic Press.
- Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science*, 304(5669), 438–41.
- Hagoort, P., & van Berkum, J. (2007). Beyond the sentence given. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 362(1481), 801–811.
- Haiman, J. (1980). Dictionaries and encyclopedias. *Lingua*, 50, 329–357.
- Haiman, J. (1982). Dictionaries and encyclopedias again. *Lingua*, 56, 353–355.
- Hare, M., Jones, M., Thomson, C., Kelly, S., & McRae, K. (2009). Activating event knowledge. *Cognition*, 111(2), 151–67.

- Haspelmath, M. (2000). The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. In *The new psychology of language: cognitive and functional approaches to language structure*, vol. II, (pp. 1–30). Mahwah, NJ: Lawrence Erlbaum.
- Haspelmath, M. (2011). The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica*, 45(1), 31–80.
- Heine, B., & Kuteva, T. (2002). On the evolution of grammatical forms. In A. Wray (Ed.) *The Transition to Language*, (pp. 376–397). Oxford: Oxford University Press.
- Hobaiter, C., & Byrne, R. (2014). The meanings of chimpanzee gestures. *Current Biology*, 24(14), 1596–1600.
- Hockett, C. F. (1960). The origin of speech. *Scientific American*, 203, 88–96.
- Hoeffler, S. H., & Smith, A. D. M. (2009). The pre-linguistic basis of grammaticalisation: A unified approach to metaphor and reanalysis. *Studies in Language*, 33(4), 886–909.
- Hopper, P. J., & Traugott, E. C. (2003). *Grammaticalization*. Cambridge: Cambridge University Press.
- Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2, 193–218.
- Hudson, R. (2007). *Language networks: The new Word Grammar*. Oxford: Oxford University Press.
- Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, 1(2), 151–195.
- Hurford, J. R. (1987). *Language and number: The emergence of a cognitive system*. Oxford: Blackwell.
- Hurford, J. R. (1989). Biological evolution of the Saussurean sign as a component of the language acquisition device. *Lingua*, 77(2), 187–222.

- Hurford, J. R. (2000). Social transmission favours linguistic generalisation. In C. Knight, M. Studdert-Kennedy, & J. R. Hurford (Eds.) *The evolutionary emergence of language: Social function and the origins of linguistic form*, (pp. 324–352). Cambridge: Cambridge University Press.
- Hurford, J. R. (2007). *The origins of meaning: Language in the light of evolution*. Oxford: Oxford University Press.
- Hurford, J. R. (2012). *The origins of grammar: Language in the light of evolution*. Oxford: Oxford University Press.
- Jackendoff, R. (1975). A system of semantic primitives. In *Proceedings of the 1975 workshop on theoretical issues in natural language processing*, vol. 77.
- Jackendoff, R. (1996). Conceptual semantics and cognitive linguistics. *Cognitive Linguistics*, 7(1), 93–129.
- Jackendoff, R. (2002). *Foundations of language*. Oxford: Oxford University Press.
- Jäger, G. (2008). The evolution of convex categories. *Linguistics and Philosophy*, 30(5), 551–564.
- Jäger, H., Baronchelli, A., Briscoe, E., Christiansen, M. H., Griffiths, T., Jäger, G., Kirby, S., Komarova, N., Richerson, P., Steels, L., & Triesch, J. (2009). What can mathematical, computational and robotic models tell us about the origins of syntax? In D. Bickerton, & E. Szathmáry (Eds.) *Biological foundations and origin of syntax*. Cambridge, MA: MIT Press.
- Jones, S. S., Smith, L. B., & Landau, B. (1991). Object properties and knowledge in early lexical learning. *Child Development*, 62(3), 499–516.
- Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, 49(1), 133–156.
- Kearns, K. (2000). *Semantics*. Basingstoke: Macmillan.
- Keil, F. C. (1994). Explanation, association, and the acquisition of word meaning. *Lingua*, 92, 169–196.

- Kelemen, D., & Bloom, P. (1994). Domain-specific knowledge in simple categorization tasks. *Psychonomic Bulletin & Review*, *1*(3), 390–395.
- Kemler Nelson, D. G. (1995). Principle-based inferences in young children's categorization: Revisiting the impact of function on the naming of artifacts. *Cognitive Development*, *10*(3), 347–380.
- Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*(1049), 1049–1054.
- Kirby, S. (2000). Syntax without natural selection: How compositionality emerges from vocabulary in a population of learners. In C. Knight (Ed.) *The evolutionary emergence of language: Social function and the origins of linguistic form*, (pp. 303–323). Cambridge: Cambridge University Press.
- Kirby, S. (2001). Spontaneous evolution of linguistic structure: an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, *5*(2), 102–110.
- Kirby, S. (2002). Natural language from artificial life. *Artificial Life*, *8*, 185–215.
- Kirby, S. (2007). The evolution of meaning-space structure through iterated learning. In C. Lyon, C. Nehaniv, & A. Cangelosi (Eds.) *Emergence of communication and language*, (pp. 253–268). Springer Verlag.
- Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(31), 10681–10686.
- Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, *28*, 108–114.
- Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (submitted). Compressibility and expressivity in the cultural evolution of linguistic structure.
- Klein, D. E., & Murphy, G. L. (2001). The representation of polysemous words. *Journal of Memory and Language*, *45*(2), 259–282.
- Klein, D. E., & Murphy, G. L. (2002). Paper has been my ruin: Conceptual relations of polysemous senses. *Journal of Memory and Language*, *47*, 548–570.

- Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99(1), 22–44.
- Lakoff, G. (1987). *Women, fire, and dangerous things*. Chicago: University of Chicago Press.
- Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.
- Landau, B., & Shipley, E. (2001). Labelling patterns and object naming. *Developmental Science*, 4(1), 109–118.
- Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, 3, 299–321.
- Langacker, R. W. (1976). Semantic representations and the linguistic relativity hypothesis. *Foundations of Language*, 14(3), 307–357.
- Laurence, S., & Margolis, E. (1999). Concepts and cognitive science. In E. Margolis, & S. Laurence (Eds.) *Concepts: Core readings*, chap. 1, (pp. 3–81). Cambridge, MA: MIT Press.
- Levinson, S. C. (2003). Language and mind: Let's get the issues straight! In D. Gentner, & S. Goldin-Meadow (Eds.) *Language in mind: Advances in the study of language and cognition*, 1979, (pp. 25–46). Cambridge, MA: MIT Press.
- Li, P., & Gleitman, L. (2002). Turning the tables: language and spatial reasoning. *Cognition*, 83(3), 265–94.
- Lin, E. L., & Murphy, G. L. (1997). Effects of background knowledge on object categorization and part detection. *Journal of Experimental Psychology: Human Perception and Performance*, 23(4), 1153–1169.
- Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2), 129–137.
- Lucas, M. (1999). Context effects in lexical access: a meta-analysis. *Memory & Cognition*, 27(3), 385–98.
- Lupyan, G., & Casasanto, D. (2014). Meaningless words promote meaningful categorization. *Language and Cognition*, (pp. 1–27).

- Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not just for talking: Redundant labels facilitate learning of novel categories. *Psychological Science*, *18*(12), 1077–1083.
- Majid, A. (2010). Words for parts of the body. In B. C. Malt, & P. Wolff (Eds.) *Words and the Mind : How words capture human experience*, (pp. 58–70).
- Majid, A., Boster, J. S., & Bowerman, M. (2008). The cross-linguistic categorization of everyday events: a study of cutting and breaking. *Cognition*, *109*(2), 235–50.
- Malotki, E. (1983). *Hopi time: A linguistic analysis of temporal concepts in the Hopi language*. Berlin: Mouton.
- Malt, B. C., & Majid, A. (2013). How thought is mapped into words. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(6), 583–597.
- Malt, B. C., Sloman, S. A., Gennari, S., Shi, M., & Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language*, *40*(2), 230–262.
- Marcus, G. F. (1993). Negative evidence in language acquisition. *Cognition*, *46*(1), 53–85.
- Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin*, *129*(4), 592–613.
- Markman, E. M. (1994). Constraints on word meaning in early language acquisition. *Lingua*, *92*, 199–227.
- Matsuki, K., Chow, T., Hare, M., Elman, J. L., Scheepers, C., & McRae, K. (2011). Event-based plausibility immediately influences on-line language comprehension. *Journal of experimental psychology. Learning, memory, and cognition*, *37*(4), 913–34.
- Matthews, C. (2009). *The emergence of categorization : Language transmission in an Iterated Learning Model using a continuous meaning space*. MSc, University of Edinburgh.
- McClelland, J. L. (2010). Emergence in cognitive science. *Topics in Cognitive Science*, *2*(4), 751–770.

- McRae, K., Ferretti, T. R., & Amyote, L. (1997). Thematic roles as verb-specific concepts. *Language and Cognitive Processes*, *12*(2/3), 137–176.
- McRae, K., & Matsuki, K. (2013). Constraint-based models of sentence processing. In R. P. G. van Gompel (Ed.) *Sentence processing*, 519, (pp. 51–77). New York: Psychology Press.
- Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207–238.
- Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology*, (pp. 89–115).
- Millikan, R. G. (1998). Language conventions made simple. *The Journal of Philosophy*, *95*(4), 161–180.
- Millikan, R. G. (2001). Purposes and cross-purposes: On the evolution of languages and language. *The Monist*, *84*(3), 392–416.
- Miyagawa, S., Ojima, S., Berwick, R. C., & Okanoya, K. (2014). The integration hypothesis of human language evolution and the nature of contemporary languages. *Frontiers in Psychology*, *5*(564).
- Monaghan, P., Mattock, K., & Walker, P. (2012). The role of sound symbolism in language learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *38*(5), 1152–64.
- Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.
- Murphy, M. L. (2010). Meaning variation: Polysemy, homonymy, and vagueness. In *Lexical meaning*, (pp. 83–107). Cambridge: Cambridge University Press.
- Newmeyer, F. J. (2002). Uniformitarian assumptions and language evolution research. (pp. 359–375). Oxford: Oxford University Press.
- Nielsen, A., & Rendall, D. (2011). The sound of round: evaluating the sound-symbolic role of consonants in the classic Takete-Maluma phenomenon. *Canadian Journal of Experimental Psychology = Revue canadienne de psychologie expérimentale*, *65*(2), 115–24.

- Nieuwland, M. S., & van Berkum, J. J. A. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience*, *18*(7), 1098–1111.
- Noh, S. M., Yan, V. X., Vendetti, M. S., Castel, A. D., & Bjork, R. A. (2014). Multi-level induction of categories: Venomous snakes hijack the learning of lower category levels. *Psychological Science*, *25*(8), 1592–1599.
- Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–61.
- Ogden, C. K., & Richards, I. A. (1923). *The meaning of meaning*. London: Routledge and Keagan Paul.
- Oliphant, M., & Batali, J. (1997). Learning and the emergence of coordinated communication. *Center for Research on Language Newsletter*, *11*.
- Owren, M. J., Dieter, J. A., Seyfarth, R. M., & Cheney, D. L. (1993). Vocalizations of rhesus (*Macaca mulatta*) and Japanese (*M. fuscata*) macaques cross-fostered between species show evidence of only limited modification. *Developmental Psychobiology*, *26*(7), 389–406.
- Paul, S. T., Kellas, G., Martin, M., & Clark, M. B. (1992). Influence of contextual features on the activation of ambiguous word meanings. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(4), 703–717.
- Peirce, J. W. (2007). PsychoPy–Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1-2), 8–13.
- Perfors, A., & Navarro, D. J. (2014). Language evolution can be shaped by the structure of the world. *Cognitive Science*, *38*(4), 775–93.
- Perniss, P., Thompson, R. L., & Vigliocco, G. (2010). Iconicity as a general property of language: evidence from spoken and signed languages. *Frontiers in psychology*, *1*(December), 227.
- Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280–91.
- Pinker, S. (1994). *The language instinct*. New York: Harper Perennial.

- Pinker, S. (1999). *Words and rules*. New York: Harper Perennial.
- Pinker, S., & Bloom, P. (1990). Natural language and natural selection. *Behavioral and Brain Sciences*, 13(4), 707–784.
- Pirog Revall, K., Tanenhaus, M. K., & Aslin, R. N. (2008). Context and spoken word recognition in a novel lexicon. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(5), 1207–1223.
- Prinz, J. J. (2002). *Furnishing the mind*. Cambridge: Cambridge University Press.
- Pustejovsky, J. (1995). *The generative lexicon*. Cambridge, MA: MIT Press.
- Pyykkönen, P., Hyönä, J., & van Gompel, R. P. G. (2010). Activating gender stereotypes during online spoken language processing: evidence from Visual World Eye Tracking. *Experimental Psychology*, 57(2), 126–33.
- 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/>
- Ramachandran, V. S., & Hubbard, E. M. (2001). Synaesthesia – A window into perception, thought and language. *Journal of Consciousness Studies*, 8(12), 3–34.
- Ramscar, M. (2010). Computing machinery and understanding. *Cognitive Science*, 34(6), 966–71.
- Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336), 846–850.
- Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: relating regularization to inductive biases through iterated learning. *Cognition*, 111(3), 317–28.
- Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, 6, 855–863.
- Regier, T. (2005). The emergence of words: Attentional learning in form and meaning. *Cognitive Science*, 29, 819–865.

- Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(4), 1436–1441.
- Rendall, D., Owren, M. J., & Ryan, M. J. (2009). What do animal signals mean? *Animal Behaviour*, *78*(2), 233–240.
- Rosch, E., & Mervis, C. B. (1975). Family resemblances: studies in the internal structure of categories. *Cognitive Psychology*, *7*, 573–605.
- Rosch, E., Mervis, C. B., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382–439.
- Ruhl, C. (1989). *On monosemy: A study in linguistic semantics*. Albany: State University of New York Press.
- Rumelhart, D. E. (1979). Some problems with the notion of literal meanings. In A. Ortony (Ed.) *Metaphor and thought*, (pp. 71–82). Cambridge: Cambridge University Press.
- run, v. (2014). *OED Online*. Oxford University Press.
- Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.
- Samuelson, L. K., & Smith, L. B. (1999). Early noun vocabularies: do ontology, category structure and syntax correspond? *Cognition*, *73*(1), 1–33.
- Sandhofer, C. M., & Smith, L. B. (2001). Why children learn color and size words so differently: Evidence from adults' learning of artificial terms. *Journal of Experimental Psychology: General*, *130*(4), 600–617.
- Sapir, E. (1921). *Language: An introduction to the study of speech*. New York: Harcourt, Brace & Company.
- Schlesinger, I. M. (1977). The role of cognitive development and linguistic input in language acquisition. *Journal of Child Language*, *4*, 153–169.
- Schouwstra, M. (2012). *Semantic Structures, Communicative Strategies and the Emergence of Language*. PhD, University of Utrecht.

- Schyns, P. G., Goldstone, R. L., & Thibaut, J. P. (1998). The development of features in object concepts. *Behavioral and Brain Sciences*, *21*(1), 1–17; discussion 17–54.
- Schyns, P. G., & Rodet, L. (1997). Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(3), 681–696.
- 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.
- Scott-Phillips, T. C., Tamariz, M., Cartmill, E. A., & Hurford, J. R. (2012). *The Evolution of Language: Proceedings of the 9th International Conference (EVOLANG9)*. Singapore: World Scientific.
- Searle, J. R. (1980). The background of meaning. In J. R. Searle, F. Kiefer, & M. Bierwisch (Eds.) *Speech act theory and pragmatics*, (pp. 221–232). Dordrecht/Boston/London: D. Reidel Publishing Company.
- Seyfarth, R. M., Cheney, D. L., & Marler, P. (1980). Monkey responses to three different alarm calls: Evidence of predator classification and semantic communication. *Science*, *210*(4471), 801–803.
- Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana, IL: University of Illinois.
- Sheskin, D. J. (2011). *Handbook of parametric and nonparametric statistical procedures*. Boca Raton, FL/London: Chapman & Hall/CRC, 5 ed.
- Silvey, C., Kirby, S., & Smith, K. (2013). Communication leads to the emergence of sub-optimal category structures. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.) *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, (pp. 1312–1317). Austin, TX: Cognitive Science Society.
- Silvey, C., Kirby, S., & Smith, K. (2014). Word meanings evolve to selectively preserve distinctions on salient dimensions. *Cognitive Science*.
- Simpson, G. B. (1981). Meaning dominance and semantic context in the processing of lexical ambiguity. *Journal of Verbal Learning and Verbal Behavior*, *20*(1), 120–136.

- Sivik, L., & Taft, C. (1994). Color naming : A mapping in the NCS of common color terms. *Scandinavian Journal of Psychology*, 35, 144–164.
- Slobin, D. I. (1996). From “Thought and language” to “Thinking for speaking”. In J. Gumperz, & S. C. Levinson (Eds.) *Rethinking linguistic relativity*, (pp. 70–96). Cambridge: Cambridge University Press.
- Smith, A. D. M. (2003a). Intelligent Meaning Creation in a Clumpy World Helps Communication. *Artificial Life*, 9, 175–190.
- Smith, A. D. M. (2008a). Protolanguage reconstructed. *Interaction Studies*, 9(1), 100–116.
- Smith, A. D. M., Schouwstra, M., de Boer, B., & Smith, K. (2010a). *The Evolution of Language: Proceedings of the 8th International Conference..* Singapore: World Scientific Press.
- Smith, K. (2002). The cultural evolution of communication in a population of neural networks. *Connection Science*, 14(1), 65–84.
- Smith, K. (2003b). Learning biases for the evolution of linguistic structure : an associative network model. In W. Banzhaf, T. Christaller, P. Dittrich, J. T. Kim, & J. Ziegler (Eds.) *Advances in Artificial Life (Proceedings of the 7th European Conference on Artificial Life)*, IIm, (pp. 517–524). Berlin: Springer Verlag.
- Smith, K. (2006). The protolanguage debate: Bridging the gap? In A. D. M. Smith, & K. Smith (Eds.) *The Evolution of Language: Proceedings of the 6th International Conference*, (pp. 315–322). Singapore: World Scientific Press.
- Smith, K. (2008b). Is a holistic protolanguage a plausible precursor to language?: A test case for a modern evolutionary linguistics. *Interaction Studies*, 9(1), 1–17.
- Smith, K., Brighton, H., & Kirby, S. (2003a). Complex systems in language evolution: the cultural emergence of compositional structure. *Advances in Complex Systems*, 6(4), 537–558.
- Smith, K., Kirby, S., & Brighton, H. (2003b). Iterated learning: a framework for the emergence of language. *Artificial Life*, 9(4), 371–86.
- Smith, K., & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition*, 116(3), 444–449.

- Smith, L. B., Colunga, E., & Yoshida, H. (2010b). Knowledge as process: contextually-cued attention and early word learning. *Cognitive Science*, *34*, 1287–314.
- Smith, L. B., & Heise, D. (1992). Perceptual similarity and conceptual structure. In B. Burns (Ed.) *Percepts, Concepts and Categories*, (pp. 233–272). Amsterdam: Elsevier B.V.
- Smith, L. B., Jones, S. S., & Landau, B. (1992). Count nouns, adjectives, and perceptual properties in children's novel word interpretations. *Developmental Psychology*, *28*(2), 273–286.
- 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), 13–9.
- Smith, L. B., & Samuelson, L. (2006). An attentional learning account of the shape bias: Reply to Cimpian and Markman (2005) and Booth, Waxman, and Huang (2005). *Developmental Psychology*, *42*(6), 1339–1343.
- Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition*. Oxford: Blackwell.
- Spivey, M. (2006). Temporal dynamics in language comprehension. In *The continuity of mind*, (pp. 169–206). Oxford: Oxford University Press.
- Spranger, M., & Pauw, S. (2009). Open-ended semantics co-evolving with spatial language. In A. D. M. Smith, M. Schouwstra, B. de Boer, & K. Smith (Eds.) *The Evolution of Language: Proceedings of the 8th International Conference, 2003*, (pp. 297–304). Singapore: World Scientific.
- Steedman, M. (2014). Evolutionary basis for human language. *Physics of Life Reviews*, *1*, 1–7.
- Steels, L. (1995). A self-organizing spatial vocabulary. *Artificial Life*, *2*(3), 319–32.
- Steels, L. (2003). Evolving grounded communication for robots. *Trends in Cognitive Sciences*, *7*(7), 308–312.
- Steels, L. (2011). Modeling the cultural evolution of language. *Physics of Life Reviews*, *8*(4), 339–56.

- Steels, L., & Belpaeme, T. (2005). Coordinating perceptually grounded categories through language : A case study for colour. *Behavioral and Brain Sciences*, 28, 469–529.
- Steels, L., Kaplan, F., McIntyre, A., & Van Looveren, J. (2002). Crucial factors in the origins of word-meaning. In A. Wray (Ed.) *The transition to language*, (pp. 252–271). Oxford: Oxford University Press.
- Tallerman, M. (2007). Did our ancestors speak a holistic protolanguage? *Lingua*, 117(3), 579–604.
- Talmy, L. (2000). *Toward a cognitive semantics*. Cambridge, MA: MIT Press.
- Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217), 1632–1634.
- Theiler, J., & Gisler, G. (1997). A contiguity-enhanced k-means clustering algorithm for unsupervised multispectral image segmentation. *Proceedings of SPIE 3159*, 3159, 108–118.
- Theisen-White, C., Kirby, S., & Oberlander, J. (2011). Integrating the horizontal and vertical cultural transmission of novel communication systems. In L. Carlson, C. Hölscher, & T. Shipley (Eds.) *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, (pp. 956–961). Austin, TX: Cognitive Science Society.
- Thompson, B., Silvey, C., Kirby, S., & Smith, K. (2014). The effect of communication on category structure. In E. A. Cartmill, S. Roberts, H. Lyn, & H. Cornish (Eds.) *The Evolution of Language: Proceedings of the 10th International Conference (EVOLANG10)*. Singapore: World Scientific Press.
- Tomasello, M. (2000). The social-pragmatic theory of word learning. *Pragmatics*, 10(4), 401–413.
- Tomasello, M. (2008). *Origins of human communication*. Cambridge, MA: MIT Press.
- Trueswell, J. C., & Tanenhaus, M. K. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language*, 33, 285–318.

- Tuggy, D. (1993). Ambiguity, polysemy, and vagueness. *Cognitive Linguistics*, 4(3), 273–290.
- Tversky, A. (1977). Features of similarity. *Psychological Review*, 84(4), 327–352.
- Tversky, B. (1989). Parts, partonomies, and taxonomies. *Developmental Psychology*, 25(6), 983–995.
- Verhoef, T., Kirby, S., & de Boer, B. (2014). Emergence of combinatorial structure and economy through iterated learning with continuous acoustic signals. *Journal of Phonetics*, 43, 57–68.
- Vinh, N. X., Epps, J., & Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11, 2837–2854.
- Vogt, P. (2005). The emergence of compositional structures in perceptually grounded language games. *Artificial Intelligence*, 167(1-2), 206–242.
- Voiklis, J., & Corter, J. E. (2012). Conventional wisdom: Negotiating conventions of reference enhances category learning. *Cognitive Science*, 36(4), 607–634.
- von Frisch, K. (1967). *The dance language and orientation of bees*. Cambridge, MA: Harvard University Press.
- Warglien, M., & Gärdenfors, P. (2011). Semantics, conceptual spaces, and the meeting of minds. *Synthese*, 190(12), 2165–2193.
- Watanabe, S., Sakamoto, J., & Wakita, M. (1995). Pigeons' discrimination of paintings by Monet and Picasso. *Journal of the Experimental Analysis of Behavior*, 63(2), 165–174.
- Wheeler, B. C., & Fischer, J. (2012). Functionally referential signals: A promising paradigm whose time has passed. *Evolutionary Anthropology: Issues, News, and Reviews*, 21(5), 195–205.
- Whorf, B. L. (1956). *Language, thought and reality*. Cambridge, MA: MIT Press.
- Wierzbicka, A. (1996). *Semantics: Primes and universals*. Oxford: Oxford University Press.

- Wilson, D. (2003). Relevance theory and lexical pragmatics. *Italian Journal of Linguistics/Rivista di Linguistica*, 15, 273–291.
- Wilson, D., & Sperber, D. (2004). Relevance theory. In L. Horn, & G. Ward (Eds.) *The handbook of pragmatics*, (pp. 607–632). Oxford: Blackwell.
- Wray, A. (1998). Protolanguage as a holistic system for social interaction. *Language and Communication*, 18, 47–67.
- Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: 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.
- Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, 114(2), 245–72.
- Xu, J., Dowman, M., & Griffiths, T. L. (2013). Cultural transmission results in convergence towards colour term universals. *Proceedings of the Royal Society B: Biological Sciences*, 280(20123073).
- Zuidema, W. (2013). Language in nature: on the evolutionary roots of a cultural phenomenon. In P.-M. Binder, & K. Smith (Eds.) *The Language Phenomenon*, The Frontiers Collection, (pp. 163–189). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Zwarts, J. (1995). The semantics of relative position. In M. Simons, & T. Galloway (Eds.) *SALT V: Proceedings from Semantics and Linguistic Theory*, (pp. 405–422). Ithaca, NY: CLC Publications (Cornell).