CONCEPTUAL COMPLEXITY AND THE EVOLUTION OF THE LEXICON

A DISSERTATION

SUBMITTED TO THE DEPARTMENT OF PSYCHOLOGY

AND THE COMMITTEE ON GRADUATE STUDIES

OF STANFORD UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Molly L. Lewis

September 2016

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Michael C. Frank) Principal Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Noah Goodman)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Ellen Markman)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Thomas Icard)

Approved for the Stanford University Committee on Graduate Studies

Preface

This thesis tells you all you need to know about...

Acknowledgments

I would like to thank...

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Your introduction here...

Chapter 2

Evidence for a complexity bias

Human languages are systems for encoding information about the world. A defining feature of a symbolic coding system is that there is no inherent mapping between the form of the code and what the code denotes (Peirce, 1931)—the color red holds no natural relationship to the meaning 'stop', the numeral *3* holds no natural relationship to three units, and in language, the word 'horse' looks or sounds nothing like the four-legged mammal it denotes. This arbitrariness of the linguistic sign has long been observed as a fundamental and universal property of natural language (Saussure, 1916, 1960; Hockett, 1960). And, despite the growing number of cases suggesting instances of non-arbitrariness in the lexicon (see Schmidtke, Conrad, & Jacobs, 2014; Dingemanse, Blasi, Lupyan, Christiansen, & Monaghan, 2015, for reviews), there is clear evidence for at least some degree of arbitrariness in language based only on the observation that different languages use different words to denote the same meaning (e.g., the word for horse in English is "horse" but is "at" in Turkish).

However, the arbitrary character of language holds only from the perspective of the analyst observing a language system from the outside; from the perspective of an individual

speaker, the goal of communication provides a strong constraint on arbitrariness. Perhaps this communicative constraint—roughly, that if my words were any different, I couldn't use them to talk to you—is why language doesn't *seem* arbitrary to us. Put another way, Saussure's (1916, 1960) insight was an insight because the form of language typically feels just right for the use to which we put it, namely talking to other people (Sutherland & Cimpian, 2015).

A rich body of theoretical work has explored communicative regularities in the use of particular forms to refer to particular types of meanings in context—the study of *pragmatics* (Grice, 1975; Horn, 1984; Clark, 1996). Broadly, this work argues that language users assume certain regularities in how speakers refer to meanings, and through these shared assumptions, the symmetry of the otherwise arbitrary character of language is broken. For example, consider a speaker who intends to refer to a particular apple on a table. Because language is *a priori* arbitrary, there are a range of ways the speaker could convey this meaning (e.g., "the apple," "the banana," "the green apple,""the green apple next to the plate," etc.), but the speaker is constrained by pragmatic pressures of the communicative context. If the listener also speaks English, the phrase "the banana" will be an unhelpful way to refer to the apple. Furthermore, if there is only one apple on the table, the phrase "the green apple" will be unnecessarily verbose given the referential context. These constraints might lead a speaker to select "the apple" as the referring expression, because it both allows the listener to correctly identify the intended referent while also minimizing effort on the part of the speaker.

In the present paper, we examine whether principles of communication influence the otherwise arbitrary mappings between words and meanings in the lexicon. This hypothesis

is motivated by a regularity first observed by Horn (1984), who noted that pragmatic language users tend to consider the effort that speakers have exerted to convey a meaning. For example, consider the utterance "Lee got the car to stop," which seems to imply an unusual state of affairs. Had the speaker wished to convey that Lee simply applied the brakes, the shorter and less exceptional "Lee stopped the car" would be a better description. The use of a longer utterance licenses the inference that there was some problem in stopping—perhaps the brakes failed—and that the situation is more complex.

We ask whether speakers reason the same way about the meanings of words, breaking the symmetry between two unknown meanings by reference to length. Specifically, we test the following hypotheses:

> *Complexity Hypothesis 1*: Speakers have a bias to believe that longer linguistic forms refer to conceptually more complex meanings.
>
> *Complexity Hypothesis 2*: Languages encode conceptually more complex meanings with longer linguistic forms.

These two hypotheses are in principle independent from one another, and we test them separately. We see them as potentially emerging together from the same interactive forces, however, and we return to this relationship in the General Discussion.

An important construct for our hypothesis is the notion of conceptual complexity. One theoretical framework for understanding this construct is through conceptual primitives (e.g., Locke, 1847). Conceptual primitives can be thought of as the building blocks of meaning, similar to the notion of geons in the study of object recognition (Biederman, 1987). Within this framework, a more complex meaning would be one with more primitives in it. In a probabilistic framework, having more units would also be correlated with having a

lower overall probability. We adopt this framework of conceptual primitives in our working definition of complexity.

Although identifying a general set of conceptual primitives might rank among the deepest challenges for cognitive science, some work has attempted this task. A body of research has sought to understand the innate conceptual primitives in young children ("core knowledge"; Kinzler & Spelke, 2007). The proposed set of concepts in this work, however, is restricted to those present only in early development (e.g., "agent"), and is therefore not suitable for the broad scope of our current project. Wierzbicka and colleagues (1996) have also sought to identify conceptual primitives, but with a more general focus. This work compares lexical systems across languages to identify common primitives. The hypothesis is that there exists universal and innate semantic primitives which are the building blocks of meaning in human language. Under this view, all meanings can be derived from a set of numerable semantic primitives and a syntax for combining them. Our work here does not directly address the character of the underlying primitives, nor whether they are universal or innate. Rather, it assumes only that such units exist for a speaker and that lexical meanings can vary in the number of their compositional primitives.

In the remainder of the Introduction, we first review prior work suggesting that communicative principles are reflected in the structure of the lexicon. We then review work related to accounts of our particular linguistic feature of interest—variability in the length of forms. Then, in the body of the paper we test the complexity hypotheses above in nine experiments and a corpus analysis.

2.0.1 Pragmatic equilibria in the lexicon

The present hypotheses are motivated by the possibility that language dynamics take place over different timescales, and these different dynamics may be causally related to each other (Christiansen & Chater, 2015; McMurray, Horst, & Samuelson, 2012; Blythe, 2015). Our two hypotheses correspond to two distinct timescales. Hypothesis 1 corresponds to the timescale of minutes in a single communicative interaction—*the pragmatic timescale*. Hypothesis 2 corresponds to the timescale of language change, which takes place over many years—*the language evolution timescale*. We consider the possibility that communicative pressures at the pragmatic timescale may, over time, influence the structure of the lexicon at the language evolution timescale. Although a complexity bias at the language evolution timescale has not been previously explored, there are a number of other cases in which pragmatic equilibria are reflected in the structure of the lexicon. Here, we describe three such cases: semantic organization, ambiguity, and one-to-one structure.

Several broad theories of pragmatics include a version of two distinct pressures on communication: the desire to minimize effort in speaking (*speaker pressure*) and the desire to be informative (*hearer pressure;* Zipf, 1936; Horn, 1984). Importantly, these two pressures trade off with each other: The optimal solution to the speaker's pressure is a single utterance that can refer to all meanings, while the optimal solution to the hearer's pressure is a longer utterance that presents no ambiguity. The utterance that emerges is argued to be an equilibrium between these two tradeoffs.[1]

At the timescale of language evolution, there are a number of cases in which these pragmatic equilibria are reflected in the lexicon. The most well-studied of these cases is the size of the semantic space denoted by a particular word. Horn (1984) argues that the hearer

[1]Note that this analysis only reflects interlocutors' *non*-aligned utilities in a communication task. Of course, both speaker and hearer also have aligned utility derived from successful communication.

has a pressure to narrow semantic space. This reflects the idea that the hearer's optimal language is one in which every possible meaning receives its own word. To understand this, consider the word "rectangle," which refers to a quadrilateral with four right angles. A special case of a "rectangle" is a case where the four sides are equal in length, which has its own special name, "square." Consequently, the term "rectangle" has been narrowed to mean a quadrilateral with four right angles, where the four sides are *not* equal. From the speaker's perspective, there is a pressure for semantic broadening. This is because the speaker's ideal language is one in which a single word can refer to a wide range of meanings. This phenomenon is exemplified by the broadening of brand names to refer to a kind of product. For example, "kleenex" is a product name for facial tissues, but has taken on the meaning of facial tissues more generally.

The opposition of these two semantic forces predicts an equilibrium in the organization of semantic space that satisfies the pressures of both speaker and hearer. A growing body of empirical work tests this prediction by examining the organization of particular semantic domains cross-linguistically (see Regier, Kemp, & Kay, 2015, for review). This work finds that languages show a large degree of similarity in how they partition semantic space for a particular domain, but also a large degree of variability. Such analyses demonstrate that the attested systems all approximate an equilibrium point between hearer and speaker pressures.

In one example of this kind of analysis, Kemp and Regier (2012) demonstrate this systematicity in the semantic domain of kinship. For each language, they developed a metric of the degree to which Horn's speaker and hearer pressures are satisfied. A language that better satisfies the hearer's pressure is one that is more complex, as measured by the description length of the system in their representational language. A language that better

satisfies the speaker's pressure is one that requires less language to describe the intended referent. To understand this, consider the word "grandmother" in English: This word is ambiguous in English because it could refer to either the maternal or paternal mother, and so identifying which mother the speaker is referring to is more costly in English than in a language that encodes this distinction lexically. They find that the set of attested languages is a subset of the range of possible languages, and this subset partitions the semantic space in a way that near optimally trades off between pragmatic pressures. This type of analysis has also been performed for the domains of color (Regier, Kay, & Khetarpal, 2007), lightness (Baddeley & Attewell, 2009), and numerosity (Xu & Regier, 2014).

A second phenomenon that is predicted by these pressures is the presence of multiple meanings associated with the same word, or lexical ambiguity. Lexical ambiguity is present in many open-class words like "bat" (a baseball instrument or a flying mammal). Lexical ambiguity is tolerated because the meaning is usually easily disambiguated by context. When the word "bat" is uttered while watching a baseball game, the mammal usage of the word is very unlikely. The presence of this type of ambiguity can be viewed as an equilibrium between the two pragmatic pressures: If the meaning of a word can be disambiguated by the referential context, then it would violate the speaker's pressure to minimize effort by keeping track of two distinct words.

Indeed, recent work by Piantadosi, Tily, and Gibson (2011b) reveals systematicity in the presence of lexical ambiguity in language. They argue that ambiguity results from a speaker based pressure to broaden the meaning of a word to include multiple possible meanings. In particular, they suggest that this pressure should lead to a systematic relationship between the presence of ambiguity and the cost of a word. According to their argument, costly words (in terms of length, frequency, or any metric of cost) that are easily understood by context

violate the speaker's principle to say no more than you must. Consequently, there should be a pressure for these meanings to get mapped on to a different, less costly word. This word may happen to already have a meaning associated with it, and so the result is multiple meanings being mapped to a single word. For example, in the case of the word "bat," a speaker could instead say "baseball bat." But, because this referent is easily disambiguated in context from the mammalian meaning, a speaker pressure should result in the use of the shorter form. This logic leads to a testable prediction: that shorter words should tend to be more ambiguous. Through corpus analyses, Piantadosi et al. (2011b) find this precise relationship between cost and ambiguity. Across English, Dutch and German, they find that shorter words are more likely to have multiple meanings.

An additional case of this lexical ambiguity is found in words that have very little context-independent meaning, known as indexicals or deictics (Frawley, 2003). These words get their meaning from the particular referential context of the utterance, and are therefore highly ambiguous from a context-independent perspective. There are many types of indexicals that are present to varying degrees across languages. Consider the temporal indexical form "tomorrow." The context-independent meaning of this word is something like "the day after the day this word is being uttered in." Critically, abstracted from any context, this word has little meaning; it is impossible to interpret without having knowledge about the day the word was uttered. This phenomenon is also present in person pronouns (e.g., "you" and "I") and spatial forms, like "here" and "there." As for lexical ambiguity, this type of ambiguity is a predicted equilibrium point from Horn's principles: If the hearer can recover the intended referent from context, the speaker would be saying more than is necessary by using an overly-specific referential term (e.g., "December 18th, 2014" vs."tomorrow"). Indexicals, therefore, provide another instance of ambiguity in lexical

systems, which may emerge as an equilibrium from the speaker's pressure to minimize effort.

Finally, the relationship between the meanings of different words can be seen as a consequence of pragmatic principles. A number of theorists have noted a bias against two words mapping onto the same meaning — that is, a bias against synonymy (Saussure, 1916, 1960; Kiparsky, 1983; Horn, 1984; Clark, 1987, 1988). This bias is an equilibrium between Horn's speaker and hearer principles. Recall that the optimal language for a speaker is one in which a single word maps to all meanings, and the optimal language for a hearer is one in which each word maps to its own meaning. Synonymy biases language toward neither of these ideals; it only results in more words for both the speaker and the hearer to keep track of. Thus, when a listener hears a speaker use a second word for an existing meaning, the hearer infers that this could not be what the speaker intended because this would violate the speaker's principle. The result is an assumption that the second word maps to a different meaning and, ultimately, a language structure that is biased against synonymy.

As one kind of evidence for this one-to-one structure in the lexicon, Horn (1984) points to a phenomenon called *blocking*. Blocking refers to cases in which an existing lexical form blocks the presence of a different, derived form with the same root. Consider the following examples:

(a) fury furious *furiosity

(b) *cury curious curiosity

In both (a) and (b), forms that would be expected, given the inflectional morphology in English, are not permitted. This is because the common root would lead to an overlap in meaning. Examples such as this provide some evidence for a one-to-one structure in language, but a one-to-one structure is a particularly difficult linguistic regularity to test

empirically. Nonetheless, it is an important regularity because it licenses certain inferences in interpreting the meaning of words. In particular, the cognitive representation of a lexical one-to-one regularity—*mutual exclusivity*—has been posited as a powerful bias in children's word learning (Markman & Wachtel, 1988; Markman, Wasow, & Hansen, 2003).

Together these phenomena—semantic organization, ambiguity, and one-to-one structure—provide three cases in which equilibria that are predicted by theories of communication at the pragmatic timescale are reflected in the structure of the lexicon at the language evolution timescale. While this similarity across timescales does not entail causality, it is suggestive of a causal relationship between the two timescales. Next, we turn to accounts at both the pragmatic and language evolution timescale for our linguistic feature of interest: length.

2.0.2 Accounts of the length of linguistic elements

Language forms vary along many dimensions, but a salient dimension is length: words and entire utterances can have dramatically different phonetic lengths. Researchers have studied this variability at both the pragmatic timescale (utterances) and the language evolution timescale (words). Our two hypotheses propose that variability at both timescales is related to the conceptual complexity of meaning. Here, we review existing work at both timescales that attempts to account for variability in language length. At the pragmatic timescale, three theories suggest that pragmatic pressures influence the length of utterances: Zipf's theory of communication, Horn's theory of communication, and Information Theory. Hypothesis 1 falls directly out of both Horn's theory of communication and Information Theory. At the language evolution timescale, two bodies of work account for word length by appealing to the predictability of the linguistic context and the conceptual 'markedness' of meaning.

While distinct from Hypothesis 2, both of these literatures are consistent with the proposal that languages use longer words to encode conceptually more complex meanings.

Zipf (1936) provided an early account of word length that appealed to a pragmatic pressure to communicate efficiently. He argued that speakers are motivated to minimize their physical effort and that this constraint could be optimally minimized by using shorter words for meanings that were used to more frequently. This leads to the prediction that there should be an inverse relationship between the length of a word and its frequency in usage—and, indeed, the empirical data suggest a robust correlation between word length and word frequency.

Others, however, have proposed different pressures at the pragmatic timescale that might influence the length of linguistic expressions. Both Horn's theory of communication and information theory predict that longer expressions should be associated with less predictable or typical meanings than their shorter counter parts. Under Horn's theory (1984), a speaker often has the choice of using two different utterances to refer to the same meaning (in truth conditional terms), and often these utterances differ in length. Horn suggests that the sentences "Lee stopped the car." and "Lee got the car to stop" have the same denotational meaning (the successful stopping of a car), though they differ in length. The claim is that this asymmetry leads to an inference on the part of the listener that the two differ in meaning.

The logic of this inference is identical to the lexical structure case above. The listener hears a speaker use a more costly phrase to express a meaning that could have been expressed in a less costly way. The listener thus infers that this other meaning could not be what the speaker intended because this would violate the speaker's principle to say no more than is necessary. Horn adds an additional layer to this argument. He suggests that not only

do these two forms differ in meaning, but that they map onto meanings in a systematic way: The longer form gets mapped on to the more unusual meaning, while the shorter form refers to the more usual meaning. Thus, in the above example, the shorter utterance would refer to a simple, average case of car stopping, while longer utterance might refer to case where something complex or unusual happened, perhaps because Lee used the emergency brake.

The source of the particular mapping between forms of different lengths and meanings is unclear. This is because in principle there are multiple equilibrium points in the mapping between form and meaning. Assuming a one-to-one constraint on the mapping, there are two possible equilibria: {short–simple, long–complex} or {short–complex, long–simple}. Both satisfy the constraint that each form gets mapped to a unique meaning. So how do speakers arrive at the {short–simple, long–complex} equilibrium? Bergen, Levy, and Goodman (in press) successfully derive this result as a consequence of the fact that {short–simple, long–complex} is a more optimal mapping for the speaker. Another possibility relies on iconicity: Hearers have a cognitive bias to map more complex sounding forms to meanings that are similarly complex.

Bergen, Goodman, and Levy (2012) provide a direct test of the length-complexity tradeoff within a communication game. In their task, partners were told that they were in an alien world with three objects and three possible utterances. In this experiment, the idea of complexity was operationalized as frequency, such that participants were instructed that each of the three different objects had three different base rate frequencies associated with them. The cost of the utterance was manipulated directly (rather than through utterance length) by assigning different monetary costs to each object. Participants' task was to communicate about one of the objects using one of the available utterances. If they successfully communicated, they received a reward. The results suggest that both the speaker and hearer

expected costlier forms to refer to less frequent meanings, consistent with Horn's predicted equilibrium between word length and meaning.

The prediction of a complexity bias at the pragmatic timescale falls more directly out of information theory. Information theory models communication as the transfer of information across a noisy channel (Shannon, 1948). Under this theory, speakers optimize information transfer (in terms of bits) by keeping the amount of information conveyed in a unit of language constant across the speech stream. A straightforward consequence of this *uniform information density* assumption is that speakers should try to lengthen unpredictable utterances. There is evidence for this prediction across multiple levels of communication. At that level of prosody, speakers tend to increase the duration of a word in cases where the word is unpredictable (highly informative) given the local (Aylett & Turk, 2004) and global (Seyfarth, 2014) linguistic context. There is also evidence for this prediction at the level of syntactic (Frank & Jaeger, 2008) and discourse predictability (Genzel & Charniak, 2002).

At the timescale of language evolution, there is some indirect evidence that this same bias is present in the lexicon. These approaches use the linguistic context of a word as a measure of the complexity of meaning. The idea is that words that are highly predictable, given the linguistic context, have more complex meanings, while words that are less predictable given the linguistic context, have less complex meanings. Piantadosi, Tily, and Gibson (2011a) measured the relationship between the predictability of a word in context and its length. Across 10 languages, these two measures were highly correlated: words that were longer were less predictable in their linguistic context on average. This result held true even controlling for the frequency of words. Additional evidence for this relationship comes from examining pairs of words that have very similar meaning, but differ

in length (e.g. "exam" vs. "examination;" Mahowald, Fedorenko, Piantadosi, & Gibson, 2012). In corpus analyses, longer forms are found to be used in less predicable linguistic contexts. They also find in a behavioral experiment that speakers are more likely to select the longer alternative in less predictive contexts. This body of work points to a systematic relationship between word length and meaning when complexity is operationalized as predictability in the linguistic context.

A related body of work has examined the relationship between length and meaning under the rubric of *markedness*, or iconicity more broadly (Jakobson, 1966). While many notions of iconicity have been discussed in the literature (Haspelmath, 2006, 2008), one version of the hypothesis is that linguistic forms often have binary morphemic contrasts and these contrasts map onto a broad difference in meaning (Greenberg, 1966). For example, consider the pair "real"–"unreal," which differ both in valence—positive vs. negative—and length (the negative form has the extra morpheme "un-"). Greenberg (1966) suggests that the difference in length is because negative meanings are conceptually more marked than their positive counterparts, and that this regularity is a linguistic universal. One explanation of this is that the set of negated things tends to be larger than the set of positive things (in principle, there are more unreal things than real things). However, a limitation of this proposal is that there is no *a priori* criteria for determining what characterizes conceptual markedness; the accounts are specific to each domain. For example, while the negation case appeals to 'number of things' as the determiner of complexity, there is no clear account of why the present form (e.g. "walk") should be less marked than the past form (e.g. "walked") or why state words (e.g. "black") should be less marked than change of state words (e.g. "blacken"). Nonetheless, this version of the markedness hypothesis suggests a relationship between linguistic length and conceptual features, similar to the complexity hypothesis.

The complexity hypothesis differs from this prior work in several ways. First, we propose conceptual complexity as a general construct that can be applied to a broad class of meanings. The hypothesis also differs in the specificity of the length metric: While markedness predicts a regularity only at the level of morphemes, the complexity hypothesis predicts a regularity at all levels of linguistic form (phonemes, syllables, morphemes). Finally, the complexity hypothesis provides an operationalization of iconicity that allows for a more direct test of the mechanism underlying systematicity between length and meaning. Haspelmath (2008) argues that the systematicity between length and meaning is not the result of a cognitive bias related to the meaning of the word, but rather due to differences in frequency of use. By providing a general definition of complexity, we are able to test for systematicity between word meaning and length, independent of frequency.

Thus, at the pragmatic timescale, there is a well-motivated prediction that less predictable meanings should be described with longer utterances. If dynamics at shorter timescales influence those at longer timescales, we might expect this same regularity to emerge in the lexicon over the course of language evolution. At the language evolution timescale, there is some indirect evidence that longer words refer to more complex meanings, but no work directly and systematically tests this prediction.

2.0.3 Our studies

The goal of our work here is to test the two complexity hypotheses given above. We present ten studies that provide support for both hypotheses: a complexity bias in individual speakers (Hypothesis 1; Experiments 1-8) and a complexity bias in natural language (Hypothesis 2; Experiments 9-10; see Table 2.1 for a summary of our studies). In Experiments 1-7, we test whether participants are biased to map a relatively long novel word onto

Experiment	Description	Complexity Hypothesis	Stimulus Type
1	Explicit complexity norms	1	artificial objects
2	Mapping task	1	artificial objects
3	Mapping task (control)	1	artificial objects
4	Explicit complexity norms	1	novel real objects
5	Mapping task	1	novel real objects
6	Mapping task (control)	1	novel real objects
7	Label production	1	novel real objects
8	Memory task to elicit RTs	1	artificial (a) and novel real (b) objects
9	English complexity norms	2	real words
10	Cross-linguistic corpus analysis	2	real words

Table 2.1: Summary of studies.

a relatively more complex object, using artificial objects (Experiments 1-3) and novel, real objects (Experiments 4-7). In Experiment 8, we explore the underlying cognitive construct of complexity in a reaction time task. In Experiment 9, we elicit complexity norms for English words and then conduct a corpus analysis of 79 additional languages (Study 10). In these studies, we operationalize the notion of conceptual complexity by manipulating it visually and also measuring it, both directly through explicit norms and indirectly through reaction time. Each approach to operationalization appeals to a broad definition of complexity where more complex meanings are assumed to have more 'parts.' In the General Discussion, we summarize the support these studies provide for our hypotheses as well as their limitations and directions for future work.

2.1 Experiment 1: Object Complexity Norms (Artificial Objects)

As a first step in exploring a complexity bias, we manipulated the complexity of objects and asked participants to infer which object a novel word refers to. Object complexity

was manipulated by varying the number of primitive parts the objects were composed of. If participants have a complexity bias, we predicted they should be more likely to map a longer novel word onto an object composed of more parts, compared to an object with fewer parts. In Experiment 1, we first conducted a norming study to verify our intuitions that the number of object parts correlated with explicit judgements of complexity. In Experiment 2, we used these normed stimuli in a simple word mapping task, revealing a complexity bias. Experiment 3 replicated Experiment 2 with randomly concatenated syllables.

2.1.1 Methods

Participants

In this and all subsequent experiments, participants were recruited on Amazon Mechanical Turk and received US $0.15-0.30 for their participation, depending on the length of the task. 60 participants completed this first experiment.

Across all experiments, some participants completed more than one experiment. The results presented here include the data from all participants, but all reported results remain reliable when excluding participants who completed more than one study. Participants were counted as a repeat participant if they completed a study using the same stimuli (e.g., completed both Experiment 1 and 2 with artificial objects).

Stimuli

As object primitives, we used "geon" shapes which are argued to be primitives in the visual system under one theory of object recognition (Biederman, 1987). We created a set of 40

Figure 2.1: Artificial objects used in Experiment 1. Each row corresponds to a complexity condition. The complexity condition is determined by the number of "geon" parts the object contains (1-5).

objects containing 1-5 geon primitives (Figure 2.1).[2]

Procedure

We presented participants 12 objects from the full stimulus set one at a time. For each object, we asked "How complicated is this object?," and participants responded using a slider scale anchored at "simple" and "complicated." Each participant saw two objects from each complexity condition, and the first two objects were images of a ball and a motherboard to anchor participants on the scale. This and all subsequent experimental paradigms can be viewed directly here: `https://mllewis.github.io/projects/RC/RCindex.html`.

[2]All stimuli, experiments, raw data and analysis code can be found at `https://github.com/mllewis/RC`. Analyses can be found at: `https://mllewis.github.io/projects/RC/RCSI.html`.

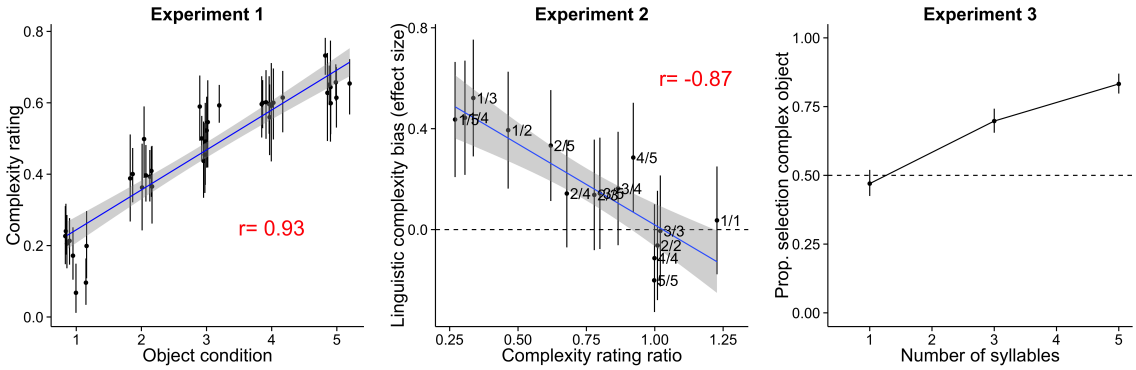

Figure 2.2: (a) The relationship between number of geons and complexity rating is plotted below. Each point corresponds to an object item (8 per condition). The x-coordinates have been jittered to avoid over-plotting. (b) Effect size (bias to select complex alternative in long vs. short word condition) as a function of the complexity rating ratio between the two object alternatives. Each point corresponds to an object condition. Conditions are labeled by the number of geons of the two alternatives. For example, the "1/5" condition corresponds to the condition in which one alternative contains 1 geon and the other contains 5 geons. (c) Proportion complex object selections as a function of the number of syllables in the target label. The dashed line reflects chance selection between the simple and complex alternatives. All errors bars reflect 95% confidence intervals, calculated via non-parametric bootstrapping in 1a and 1c, and parametrically in 1b.

2.1.2 Results and Discussion

Number of object parts was highly correlated with explicit complexity judgment ($r = .93$, $p < .0001$; $M = .47$, $SD = .18$): Objects with more parts tend to be rated as more complex.[3] Figure 2.2a shows the mean complexity rating for each of the 40 objects as a function of their complexity condition. This finding suggests that we can use manipulations of visual complexity as a proxy for manipulations of conceptual complexity.

[3]We are interested in the relationship between measurements (specifically, word length and complexity), rather than participant-wise variability. We therefore conduct most of our analyses on item means. All correlations reported are at the item level, with the exception of Experiments 2 and 5 where we report the correlation across effect sizes. In Experiments 3, 6 and 7, we use linear mixed effect models due to the repeated-measure design in these experiments.

2.2 Experiment 2: Mapping Task (Artificial Objects)

2.2.1 Methods

Participants

750 participants completed the experiment.

Stimuli

The referent stimuli were the set of 40 objects normed in Experiment 1. The linguistic stimuli were novel words either 2 or 4 syllables long (e.g., "bugorn" and "tupabugorn"). There were 8 items of each syllable length.

Procedure

We presented participants with a novel word and two possible objects as referents, and asked them to select which object the word named ("Imagine you just heard someone say *bugorn*. Which object do you think *bugorn* refers to? Choose an object by clicking the button below it.").

Within participants, we manipulated word length and the relative complexity of the referent alternatives. We tested every unique combination of object complexities (1 vs. 2 geons, 1 vs. 3 geons, 1 vs. 4 geons, etc.), giving rise to 15 conditions in total. Each participant completed 4 short and 4 long trials in a random order, where each word was randomly associated with one of the complexity conditions. No participant saw the same complexity condition twice and no word or object was repeated across trials.

2.2.2 Results and Discussion

Across conditions, the more complex object was more likely to be judged the referent of the longer word. For each object condition (e.g., 1 vs. 2 geons), we calculated the effect size for participants' complexity bias—the degree to which the complex object was more likely to be chosen as the referent of a long word, compared to the short word. Effect sizes were calculated using the log odds ratio (Sánchez-Meca, Marín-Martínez, & Chacón-Moscoso, 2003). Effect size was highly correlated with the ratio of object complexities: The greater the mismatch in object complexity, the more the longer word was paired with the more complex object ($r = -.87$, $p < .0001$). This experiment thus provides initial evidence for a complexity bias in the lexicon: Given an artificial word and two objects of differing visual complexity, participants are more likely to map a longer word onto a more complex referent, relative to a shorter word.

2.3 Experiment 3: Control Mapping Task (Artificial Objects)

One limitation of Experiment 2 is that it uses a small set of words as the linguistic stimuli (8 short and 8 long), making it possible that idiosyncratic properties of the words could be driving the observed complexity bias. In Experiment 3, we sought to test this possibility by using words composed of randomly concatenated syllables rather than items selected from a small list of words. The design was identical to Experiment 2, except that we tested only the most extreme complexity condition, the "1/5" condition.

2.3.1 Methods

Participants

200 participants completed the experiment.

Stimuli

The referent stimuli were the geon objects composed of either 1 or 5 geons. The novel words were created by randomly concatenating 1, 3, or 5 consonant-vowel syllables. The last syllable of all words ended in a consonant to better approximate the phonology of English (e.g., "nur," "nobimup," "gugotobanid").

Procedure

Participants completed six forced-choice trials identical to Experiment 1b. We tested only the "1/5" complexity condition (1-geon object vs. 5-geon object). Word length was manipulated within-participant such that each participant completed 2 trials for each of the three possible word lengths (1, 3, or 5 syllables).

2.3.2 Results and Discussion

To examine the effect of length on referent selection, we constructed a generalized linear mixed-effect modeling predicting referent selection with word length. We included random by-participant intercepts and slopes. Replicating the "1/5" condition in Experiment 2, we found that participants were more likely to select a five geon object compared to a single geon object as the number of syllables in the word increased ($\beta = -.60$, $z = -8.63$, $p < .0001$). This finding suggests that the complexity bias observed in Experiment 2 is unlikely

to be due to the particular set of words we selected.

2.4 Experiment 4: Object Complexity Norms (Novel Objects)

Experiments 1-3 provide evidence for a complexity bias using artificial objects. The complexity manipulation in these experiments was highly transparent, however, making it possible that task demands influenced the effect. We next asked whether this bias extended to more naturalistic objects where the variability in complexity might be less obvious to participants. We conducted the same set of 3 experiments as above using a sample of real objects without canonical labels. We find that the complexity bias observed with artificial geon objects extends to naturalistic objects.

2.4.1 Methods

Participants

We recruited two samples of 60 participants to complete Experiment 4.

Stimuli

We collected a set of 60 objects that were real objects but that we judged not to have canonical labels associated with them (Figure 2.3).

Procedure

The procedure was identical to Experiment 1.

Figure 2.3: Novel real objects used in Experiments 4-6: Naturalistic objects without canonical labels. Each row corresponds to a quintile determined by the explicit complexity judgments obtained in Experiment 4 (top: least complex; bottom: most complex).

2.4.2 Results and Discussion

Complexity judgments were highly reliable across two independent samples ($r = .93, p < .0001$; $M_1 = .49$, $SD_1 = .18, M_2 = .44$, $SD_2 = .18$; mean difference = .07). Figure 2.4a shows the relationship between the complexity judgment for each item across the two samples of participants. Figure 2.3 shows all 60 objects sorted by their mean complexity rating.

2.5 Experiment 5: Mapping Task (Novel Real Objects)

2.5.1 Methods

Participants

1500 participants completed the experiment.

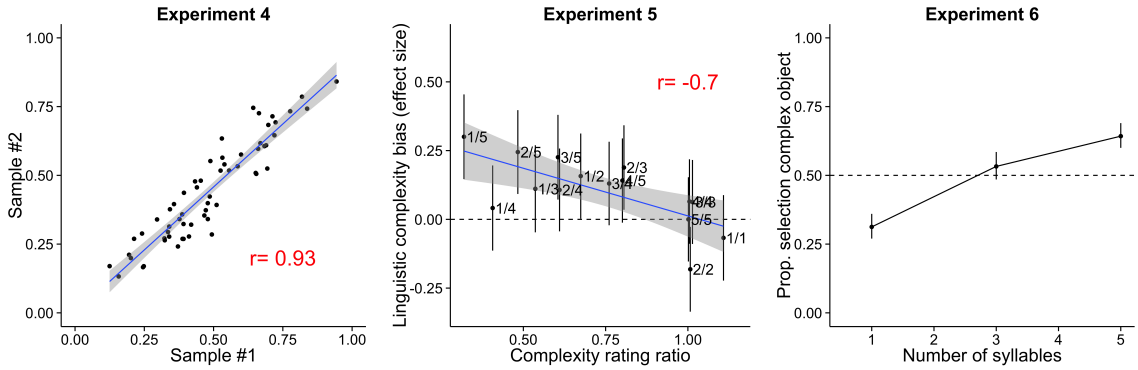

Figure 2.4: (a) The correlation between the two samples of complexity norms. Each point corresponds to an object ($n = 60$). (b) Effect size (bias to select complex alternative in long vs. short word condition) as a function of the complexity rating ratio between the two object alternatives. Each point corresponds to an object condition. Conditions are labeled by the complexity norm quintile of the two alternatives. (c) The proportion of complex object selections as a function of number of syllables. The dashed line reflects chance selection between the simple and complex alternatives. All errors bars reflect 95% confidence intervals, calculated via non-parametric bootstrapping in 4 and 6, and parametrically in 5.

Stimuli

The linguistic stimuli were identical to Experiment 2. The object stimuli were the 60 naturalistic objects normed in Experiment 2. Five complexity conditions were determined by dividing the objects into quintiles based on the norms.

Procedure

The procedure was identical to Experiment 2, except for the use of naturalistic rather than artificial geon objects.

2.5.2 Results and Discussion

As with the artificial objects, effect size was negatively correlated with the complexity rating ratio between the referent alternatives ($r = .70, p < .005$; Fig. 2.4b). This suggests that the complexity bias observed with artificial objects extends to more naturalistic objects, consistent with the proposal that a complexity bias is a characteristic of natural language more generally.

The effect size in Experiment 5 is smaller than in Experiment 2, however. This difference may be due to the fact that some of the effect in Experiment 2 was due to task demands associated with the transparent complexity manipulation. Nonetheless, Experiment 5 reveals a complexity bias with naturalistic objects.

2.6 Experiment 6: Control Mapping Task (Novel Objects)

As with the artificial objects, we sought to control for the possibility that the results from the mapping task were due to our particular linguistic items. Thus, we conducted a control experiment analogous to Experiment 3 using randomly concatenated syllables.

2.6.1 Methods

Participants

200 participants completed the experiment.

Stimuli

The objects were 12 objects from the first and fifth quintile of complexity norms. The linguistic stimuli were constructed as in Experiment 3.

Procedure

The procedure was identical to Experiment 3, except for the different object stimuli.

2.6.2 Results and Discussion

We fit the same model as in Experiment 3, predicting response value with length using a generalized liner mixed-effect model. A model with random by-participant slopes and intercepts failed to converge, and so the final model included only random by-participant intercepts. Participants were more likely to select an object from the fifth quintile as opposed to the first quintile when the novel word contained more syllables ($\beta = -.35$, $z = -.91$, $p < .0001$; Fig. 2.4c). This pattern replicates the complexity bias seen in Experiment 5 with randomly concatenated syllables.

In the present experiment, participants were overall less likely to select the complex object, compared to the same experiment with artificial objects (consider the overall higher level of complex-object judgments in Experiment 5). This finding may be due to the fact that some of the simple artificial objects in Experiment 3 are associated with canonical labels (e.g., the sphere single-geon object may have evoked the label "ball."). Perhaps this feature of hte stimuli might have lead participants to appeal to mutual exclusivity in their object selections by selecting an object they do not already have a name for—in this case, the more complex object (Markman & Wachtel, 1988). Alternatively, the novel artificial objects could be overall less complex than the geon objects. Regardless of this shift, however, the critical finding is that we replicate the complexity bias with random syllables in both Experiments 3 and 6.

2.7 Experiment 7: Label Production Task (Novel Objects)

The previous set of experiments provides evidence for a complexity bias in a comprehension task with novel words. One limitation of this design, however, is that participants may have been influenced by task demands associated with making a forced choice between two contrasting alternatives. In Experiment 7, we sought to minimize these demands by presenting participants with an object and asking them to produce a novel label to refer to it. Consistent with a complexity bias, we find that participants produce longer labels for more complex objects.

2.7.1 Methods

Participants

Fifty-nine participants completed the experiment.

Stimuli

The objects were drawn from the set of 60 naturalistic objects used in Experiments 4-6

Procedure

In each trial, we presented a single object and asked participants to generate a novel single-word label to refer to it. The instructions read:

> What do you think this object is called? For example, someone might call it a *tupa* or a *pakuwugnum*. In the box below, please make up your own name for the object. Your name should only be one word. It should not be a real English word.

Each participant completed 10 trials—five objects from the bottom and top complexity norm quantiles each. Order of objects was randomized.

2.7.2 Results and Discussion

There were 26 productions (4%) that included more than one word. These productions were excluded. Length was measured in terms of log number of characters.

Participants produced novel coinages that varied in length (e.g., "keyo," "plattle," "scrupula," "frillobite"). Critically, productions tended to be longer for the top quartile of objects ($M = 7.13$, $SD = 1.81$ characters) compared to the bottom quartile ($M = 6.60$, $SD = 1.78$ characters). To test the reliability of this difference, we fit a linear mixed-effect model predicting log length in terms of number of characters with complexity norm as a fixed effect. The random effect structure included by-participant intercepts and slopes. There was a reliable effect of complexity norms, suggesting that productions tended to be longer for more complex objects ($\beta = .19$, $t = 4.36$). This experiment provides strong evidence for a productive complexity bias: Even with minimal task demands, participants prefer to use longer words to refer to more complex objects.

2.8 Experiments 8a and 8b: Complexity as a Cognitive Construct

Experiments 1–7 suggest that participants have a productive complexity bias when complexity is operationalized in terms of explicit norms. In Experiment 8, we try to more directly examine the cognitive correlates of conceptual complexity. We reasoned that if complexity is related to a basic cognitive process, we should be able to measure it using an

implicit task, not just via explicit ratings.

To measure complexity implicitly, we adopt a measure from the visual processing literature: reaction time. In this literature, the amount of information in a stimulus is argued to be monotonically related to the amount of time needed to respond to that stimulus. Hyman (1953) demonstrated this using a task in which participants were asked to indicate which light was illuminated from a set of bulbs. Two factors were manipulated to vary the amount of information in each bulb: the number of bulb alternatives and the frequency of each bulb illuminating. They found that the reaction time for responding to an illuminated bulb was linearly related to the amount of information in that bulb. More recently, Alvarez and Cavanagh (2004) used a reaction time measure—search rate—to quantify the amount of information in a varied set of visual stimuli. They found that the search rate of a visual stimulus was monotonically related to the memory capacity for that stimulus. Finally, in the domain of sentence processing, reaction time has been directly correlated with measures of surprisal of a word in its linguistic context (Demberg & Keller, 2008; Levy, 2008). Together, these results suggest that reaction time may be a behavioral correlate of the amount of information, or complexity, of a stimulus.

To collect an implicit measure of complexity for our objects, we measured participants' study time of objects in a memory task. Each participant studied half of the objects in the stimulus set, one at a time, and then made old/new judgments for the entire set. Critically, the study phase was self-paced, such that participants were allowed to study each object for as much time as they wanted. This study time provided an implicit measure of complexity. For both the artificial (Experiment 8a) and naturalistic (Experiment 8b) objects, we found that participants tended to study objects longer when they were rated as more complex.

2.8.1 Methods

Participants

750 participants completed the task. 250 participants were tested with artificial objects (Experiment 8a) and 500 were tested with novel real objects (Experiment 8b).

Stimuli

The study objects were the set of 40 artificial objects (Experiment 8a) and 60 novel real objects (Experiment 8b).

Procedure

Participants were told they were going to view some objects and their memory of those exact objects would later be tested. In the study phase, participants were presented with half of the full stimulus set one at a time (20 artificial objects and 30 novel real objects) and allowed to click a "next" button when they were done studying each object. After the training phase, we presented participants with each object in the full stimulus set (40 artificial objects and 60 novel real objects), and asked "Have you seen this object before?." Participants responded by clicking a "yes" or "no" button.

2.8.2 Results and Discussion

Experiment 8a: Artificial objects

We excluded subjects who performed at or below chance on the memory task (20 or fewer correct out of 40). A response was counted as correct if it was a correct rejection or a hit. This excluded 9 participants (4%). With these participants excluded, the mean correct

was 72%. Participants were also excluded based on study times. We transformed the time into log space, and excluded responses that were 2 standard deviations above or below the mean. This excluded 4% of responses (final sample: $M = 2.02$, $SD = 1.37$ s).

Next, we examined study times for each object in this ($M = 1.89$, $SD = .28$ s). Study times were highly correlated with the number of geons in each object ($r = .93$, $p < .0001$): objects that contained more geons tended to be studied longer. Study times were also highly correlated with the explicit complexity norms ($r = .89$, $p < .0001$): objects that were rated as more complex tended to be studied longer.

The critical question was whether mean study times for an object were related to the bias to assign a long or short word to that object. To explore this question, we reanalyzed the data from Experiment 2 in terms of study times instead of explicit complexity norms. The ratio of study times for the two object alternatives was correlated with the bias to choose a longer label ($r = .82$, $p < .001$; Fig. 2.5a): Relatively longer study times predicted longer labels.

Experiment 8b: Novel real objects

We excluded six (1%) participants who performed at or below chance on the memory task (30 or fewer correct out of 60). A response was counted as correct if it was a correct rejection or a hit. With these participants excluded, the mean correct was 84%. Participants were also excluded based on study times, using the same criteria as in Experiment 8a, leading to the exclusion of 4% of responses (final sample: $M = 2.01$, $SD = 1.45$ s).

We next examined study times by object ($M = 1.92$, $SD = .18$ s). Study times were highly correlated with explicit complexity norms for each object. Like for the geons, objects that were rated as more complex were studied longer ($r = .54$, $p < .0001$). This

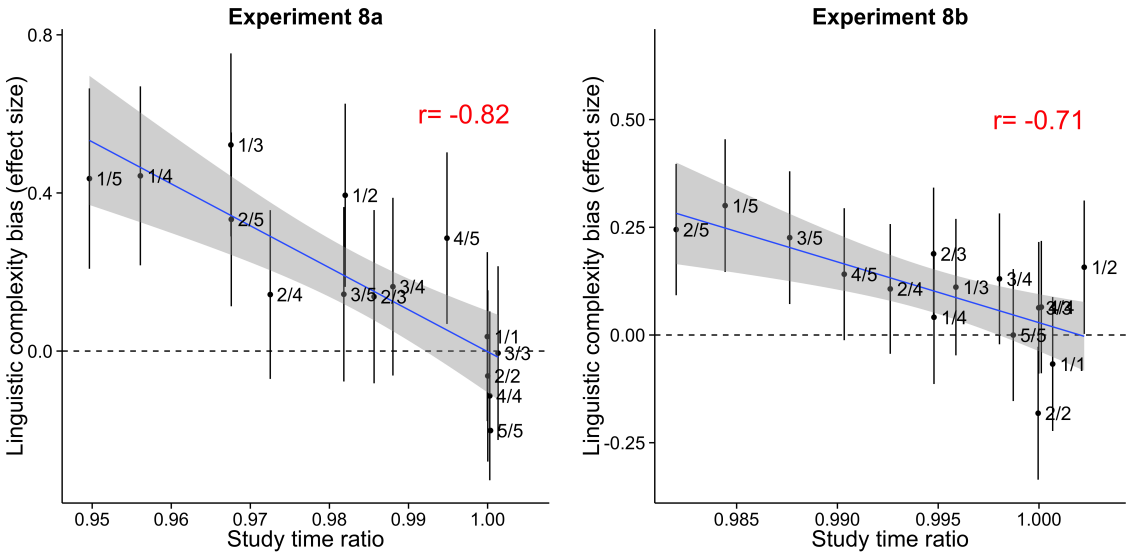

Figure 2.5: Effect sizes in Experiments 2 and 4 replotted in terms of study times collected in Experiment 8. Objects that are studied relatively longer are more likely to be assigned a longer label, relative to a shorter label. Error bars show 95% confidence intervals.

correlation was somewhat smaller than for the geons ($r = .89$), which may be due to the fact that overall variance in study times was smaller for the real objects ($SD = .18$), relative to the geons ($SD = .28$). Critically, by reanalyzing data from Experiment 4 in terms of study times, we find that the ratio of study times for the two objects was correlated with the bias to choose a longer label ($r = .71$, $p < .005$; Fig. 2.5b).

Together, these findings suggest that label judgments are supported by basic cognitive processes related to the complexity or information content of a stimulus. More broadly, Experiments 1-8 point to a complexity bias in interpreting novel labels: Words that are longer tend to be associated with meanings that are more complex, as reflected in both explicit and implicit measures.

2.9 Experiment 9: Complexity Bias in Natural Language

Experiments 1–8 revealed a productive complexity bias in the case of novel words (Hypothesis 1). Next we ask whether this bias extends to natural language (Hypothesis 2). In Experiment 9, we collected explicit complexity judgments on the meaning of 499 English words in a rating procedure similar to Experiments 1 and 4 above. Consistent with a complexity bias, we find that complexity ratings are highly correlated with word length in English: Words with meanings that are rated as more complex tend to be longer.

To measure conceptual complexity in natural language, we adopt a rating scale approach similar to that used in previous work (e.g., Wilson, 1988) to quantify other aspects meaning, like how perceptible a referent is (concreteness) and how much experience speakers tend to have with a referent (familiarity). In this work, participants are presented with a 5- or 7- point Likert scale anchored at both ends of the target dimension and asked to make an explicit judgment about a word's meaning. A limitation of this approach is that it requires that all participants conceptualize the dimension of interest in a similar way. Nonetheless, previous work has shown these measures to be reliable and so we adopt them here to quantify conceptual complexity.

2.9.1 Methods

Participants

246 participants completed the norming procedure.

Stimuli

We selected 499 English words from the MRC Psycholinguistic Database (Wilson, 1988) that were broadly distributed in their length and were relatively high frequency. This database includes norms for three other psycholinguistic variables: concreteness, familiarity, and imageability. This selection of items allowed us to compare our complexity norms to previously measured psycholinguistic variables that are intuitively related to complexity.

Procedure

Participants were first presented with instructions describing the norming task:

> In this experiment, you will be asked to decide how complex the meaning of a word is. A word's meaning is simple if it is easy to understand and has few parts. An example of a simple meaning is "brick." A word's meaning is complex if it is difficult to understand and has many parts. An example of a more complex meaning is "engine."

For each word, we then asked "How complex is the meaning of this word?," and participants indicated their response on a 7-pt Likert scale anchored at "simple" and "complex." The first two words were always "ball" and "motherboard" to anchor participants on the scale. Each participant rated a sample of 30 words English words. After the 17th word, participants were asked to complete a simple math problem to ensure they were engaged in the task.

2.9.2 Results and Discussion

We first examined word length in our samples of words, using three different metrics of word length: phonemes, syllables, and morphemes. Measures of phonemes and syllables were taken from the MRC corpus (Wilson, 1988) and measures of morphemes were taken from CELEX2 database (Baayen, Piepenbrock, & Gulikers, 1995). All three metrics were highly correlated with each other (phonemes and syllables: $r = .89$; phonemes and morphemes: $r = .65$; morphemes and syllables: $r = .67$). All three metrics were also highly correlated with number of characters, the unit of length with use in the cross-linguistic corpus analysis below (phonemes: $r = .92$; morphemes: $r = .69$; syllables: $r = .87$).

Given these measures of word length, we next considered how length related to judgments of meaning complexity. We excluded 6 participants (2%) who missed the math problem, our attentional check. Critically, we found that complexity ratings ($M = 3.36$, $SD = 1.14$) were positively correlated with word length, measured in phonemes, syllables, and morphemes ($r_{phonemes} = .67$, $r_{syllables} = .63$, $r_{morphemes} = .43$, all ps $< .0001$, Fig. 2.6).[4] This relationship held for the subset of only open class words ($n = 438$; $r_{phonemes} = .65$, $r_{syllables} = .63$, $r_{morphemes} = .42$, all ps $< .0001$). Word class was coded by the authors.

This result points to a relationship between conceptual complexity and word length, but to interpret this relationship, it is important to also control for other known correlates of word length and complexity. Linguistic predictability is highly correlated with word length, operationalized via simple frequency (Zipf, 1936) or using a language model (Piantadosi et al., 2011b). We estimated word frequency from a corpus of transcripts of American English movies (Subtlex-us database; Brysbaert & New, 2009). Importantly, the regularity we describe—a relationship between conceptual complexity and word length—holds even

[4] All norms can be found here: `https://github.com/mllewis/RC/blob/master/data/norms/`.

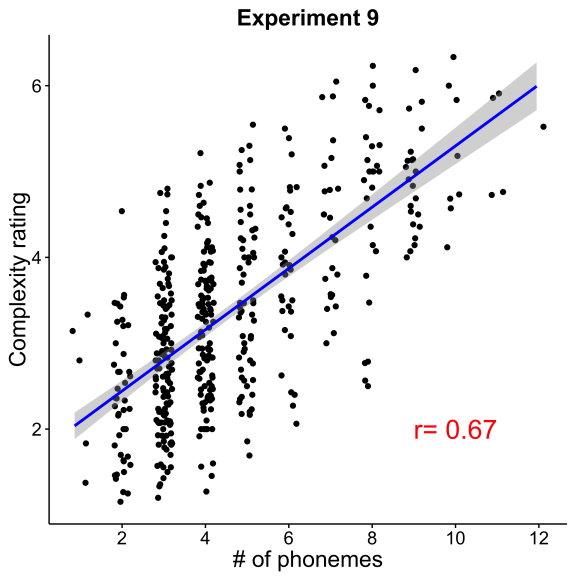

Figure 2.6: Complexity norms collected in Experiment 9 as a function of word length in terms of number of phonemes. Words rated as more complex tend to be longer. Error bars show bootstrapped 95% confidence intervals.

when controlling for frequency. In English, the correlation was only slightly reduced when controlling for log frequency ($r = .57$, $p < .0001$).

We also looked at the relationship between length and complexity controlling for the average predictability of a word in a linguistic context (its surprisal). As discussed in the Introduction, recent work suggests that surprisal may be stronger correlate of length than frequency (Piantadosi et al., 2011a). We included bigram surprisal values for our set of 499 words calculated from the British National Corpus (Clear, 1993).[5] Surprisal was correlated with complexity ($r = .29$, $p < .0001$), but the correlation between length in phonemes and complexity remained reliable after partialing out surprisal ($r = .62$, $p < .0001$). In an additive linear model predicting word length (phonemes) with complexity, frequency, and surprisal, complexity and surprisal were reliable predictors of length ($\beta = 1.11$, $t = 17.22$,

[5]We thank Steve Piantadosi for sharing this data with us.

	Estimate	Std. Error	t-value	p
(Intercept)	7.5020	0.2061	36.40	$<.001$
complexity	0.2429	0.0116	20.86	$<.001$
concreteness	-0.0033	0.0004	-9.16	$<.001$
imageability	-0.0003	0.0004	-0.81	0.42
familiarity	0.0024	0.0005	4.80	$<.001$
log frequency	-1.1556	0.0332	-34.80	$<.001$

Table 2.2: Model parameters for linear regression predicting word length in terms of semantic variables and word frequency.

$p < .0001$; $\beta = .66$, $t = 2.3$, $p = .02$), but frequency was not ($\beta = .04$, $t = .39$, $p = .70$).

Complexity is reliably correlated with concreteness, familiarity, and imageability (concreteness: $r = -.27$; familiarity: $r = -.43$; imageability: $r = -.21$). Nonetheless, the relationship between word length and complexity remained reliable controlling for these factors. We created an additive linear model predicting word length in terms of phonemes with complexity, controlling for concreteness, imageability, familiarity, and frequency. Model parameters are presented in Table 2.2. This pattern held for the other two metrics of word length (morphemes and syllables).

This result extends beyond the findings of previous work on markedness. Although this difference in the complexity of morphological structure could in principle contribute to conceptual complexity judgments, it does not explain the pattern in our data. The correlations we observed hold for words with no obvious derivational morphology (CELEX2 monomorphemes; Baayen et al., 1995, $n = 387$; $r_{phonemes} = .53$, $r_{syllables} = .47$, all ps $< .0001$).

Finally, languages also show phonological iconicity effects, such that semantic features (Maurer, Pathman, & Mondloch, 2006) and even particular form classes (Farmer, Christiansen, & Monaghan, 2006) are marked by particular sound patterns. However, the type of iconicity explored here is broader—a systematic relationship between abstract

measures of complexity and amount of verbal or orthographic effort. Specific iconic hypotheses that posit a parallel between an object's parts and the number of phonemes, morphemes, or syllables in its label do not account for the patterns in the English lexicon: The length-complexity correlation holds even more strongly for words that are not object labels ($n = 336$; $r_{phonemes} = .73$, $p < .001$), compared to object labels ($n = 163$; $r_{phonemes} = .44$,$p < .001$), whose part structure is presumably much less obvious. If true, this suggests the effect sizes in Experiments 1-8 may be conservative estimates of the bias since all referents in these experiments were concrete objects.

While correlational nature of this study makes inferences about causality tentative—complex meanings may be assigned longer words, or words that are longer may be rated as more complex—this study nonetheless points to a robust relationship between word length and conceptual complexity in English.

2.10 Study 10: Cross-Linguistic Corpus Analysis

If the complexity bias relies on a universal cognitive process, it should generalize to lexicons beyond English. We explored this prediction in 79 additional languages though a corpus analysis, and found a complexity bias in every language we examined.

2.10.1 Methods and Results

We translated all 499 words from Experiment 9 into 79 languages using Google translate (retrieved March 2014). The set of languages was the full set available in Google translate. Words that were translated as English words were removed from the data set. We also removed words that were translated into a script that was different from the target language

(e.g., an English word listed for Japanese).

Native speakers evaluated the accuracy of these translations for 12 of the 79 languages. Native speakers were told to look at the translations provided by Google, and in cases where the translation was bad or not given, provide a "better translation." Translations were not marked as inaccurate if the translation was missing. Across the 12 languages, there was .92 native speaker agreement with the Google translations across all 499 words.

To test for a complex bias, we calculated the length of each word in each of the 79 languages using number of unicode characters as our unit of length (to allow comparison between languages for which no phonetic dictionary was available). For each language, we calculated the correlation between word length in terms of number of characters and mean complexity rating. All 79 languages showed a positive correlation between length and complexity ratings. The grand mean correlation across languages was .34 (r = .37, for checked languages only).

This relationship between word length and complexity remained reliable in a number of control analyses. There was a reliable correlation between length and complexity for the subset of English monomorphic words (grand mean r = .23) and open class words (grand mean r = .30). It also held partialling out frequency (grand mean r = .22).

Finally, it is possible that the cross-linguistic regularity is due primarily to a genetic relationship between languages or language contact (Jaeger, Graff, Croft, & Pontillo, 2011). Such a finding would suggest that the bias may be an idiosyncratic property of a few languages, rather than a broad generalization of human languages. To test this possibility, we used data from the World Atlas of Language Structures database (WALS; Haspelmath, Dryer, Gil, & Comrie, 2005). We included language family as a control for genetic relationships and native country as a control for language contact. Data was available for 68 of

our 80 languages in this dataset. Within these languages, there were 16 language families and 49 countries represented.

We constructed a mixed effect model predicting word length in terms of number of characters with complexity ratings and log frequency as fixed effects. The random-effect structure included language family as both random slopes and intercepts.[6] The model showed a reliable effect of complexity on length ($\beta = .70$, $t = 3.59$), suggesting that the complexity bias is present in a wide range of languages.

2.10.2 Discussion

This corpus analysis suggests that the complexity bias found in natural language (Experiment 9) generalizes to a broad range of other languages. A notable result from these analyses is that English appears to have the largest complexity bias of the languages examined. One possible explanation is that, because our complexity norms were elicited for English words, our measure of conceptual complexity was most accurate for English words, and thus the complexity bias was largest for English. If true, then the cross-linguistic estimates of complexity bias obtained in the present analyses would be conservative estimates of a larger bias.

[6]The model specification was as follows: `word length` $\sim$ `complexity + log frequency + (1 + complexity + log frequency | language family) + (1 + complexity + log frequency | native country)`. This structure was the maximal random effect structure that allowed the model to converge.

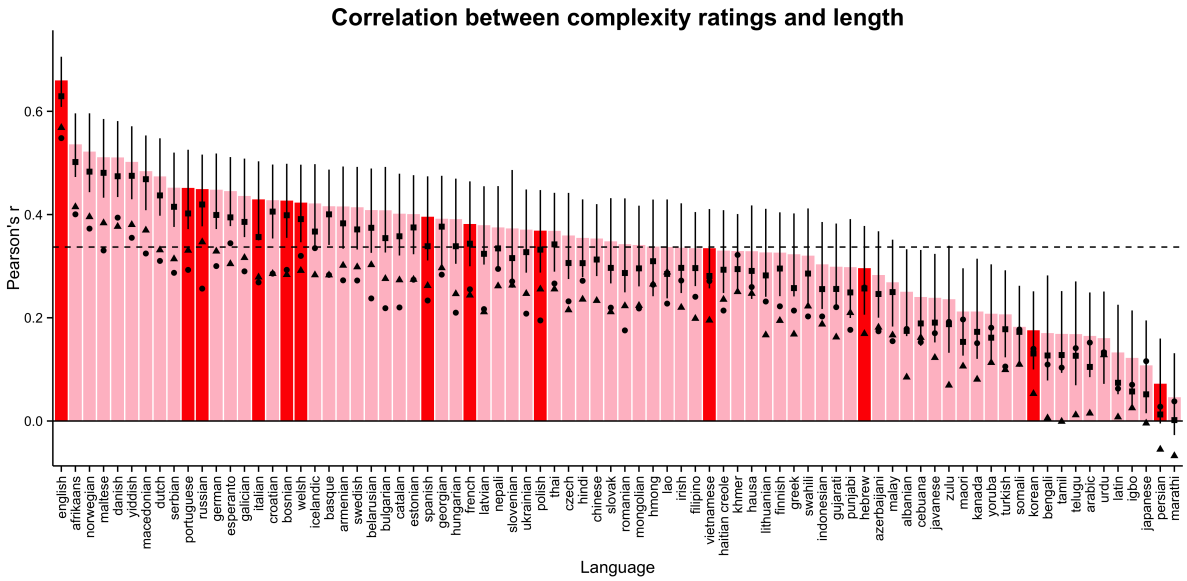

Figure 2.7: Correlation coefficient (Pearson's *r*) between length in unicode characters and conceptual complexity rating (obtained in Experiment 9). Dark red bars indicate languages for which translations were checked by native speakers; all other bars show translations obtained via Google Translate. The dashed line indicates the grand mean correlation across languages. Triangles indicate the correlation between complexity and length, partialling out log spoken frequency in English. Circles indicate the correlation between complexity and length for the subset of words that are monomorphemic in English. Squares indicate the correlation between complexity and length for the subset of open class words. Error bars show 95% confidence intervals obtained via non-parametric bootstrap.

2.11 General Discussion

We began with two observations—the presence of many pragmatic equilibria reflected in the structure of the lexicon, and the fact that several theories of pragmatics predict a tradeoff between length and complexity. The goal of our work was to explore whether a tradeoff between length and complexity is present in words—namely, a bias for longer words to refer to more conceptually more complex meanings. We explored this bias at two timescales. At the pragmatic timescale, we asked whether participants would be biased to assign a relatively long novel word to a conceptually more complex referent (Hypothesis 1). At the language evolution timescale, we asked whether languages tended to encoded conceptually more complex meanings with longer forms (Hypothesis 2). We found support for both hypotheses.

Experiments 1–7 suggest that when conceptual complexity is operationalized via visual complexity, participants are biased to assign novel words to more complex referents. This pattern holds true for both artificial objects where visual complexity was directly manipulated, as well as for naturalistic objects where we measured visual complexity and analyzed it correlationally. We also found this pattern across both comprehension and production tasks, suggesting this bias was not merely the result of task demands. Experiment 8 reveals that visual complexity is highly correlated with an implicit measure—study time—and this measure predicts the bias to assign an object a long or a short word. Finally, Experiment 9 suggests that explicit measures of conceptual complexity in English are highly correlated with word length in English, and the corpus analysis reveals a correlation between English complexity norms and word lengths in a diverse set of languages.

These studies reveal a regularity in language that appears to be productive and true cross-linguistically. The observed bias is highly general, both in terms of the unit of length

(phonemes, morphemes, and syllables) as well as the characterization of semantics. This work contributes an important extension to the previous work on markedness. Previous work on markedness described relationships between conceptual features and word length that were post-hoc and domain specific. Our work suggests that conceptual complexity may be a unifying framework for thinking about variability in conceptual space across semantic domains. In our work here, we begin to directly address the cognitive construct underlying conceptual complexity by revealing a strong relationship between explicit measures of complexity and the implicit measure of reaction time.

While the broad nature of the regularity we describe is a strength, our work here leaves a number of open questions. Additional research needs to be done to better understand what conceptual complexity is and what constructs our measures here describe. Our reaction time results suggest that, whatever conceptual complexity is, it is related to basic cognitive processes. But our work does not provide any insight into what the conceptual primitives are such that some meanings are more conceptually complex than others. We turn to this issue in Chapter 3.

A second limitation of our work is that we are not able to provide an account of why word lengths can change over time for the same meaning (e.g., "television" becomes "TV" or "cellular phone" becomes "cell"). The answer to this question may be related to the question of conceptual complexity. One possibility is that the conceptual complexity of a word's meaning may reduce over time, and language reflects this change by shortening the length of the word. Another possibility is that this reduction is the result of another pressure on language change: word frequency. Under this hypothesis, as a word become more frequent, it becomes shorter (Zipf, 1936), and this pressure is independent of the complexity bias. So perhaps such shortenings are unrelated to the phenomenon we describe

here.

Finally, our interpretation of this work is limited by the fact that all participants were speakers of English. A complexity bias could in principle be idiosyncratic to English. The results from our experiments with novel words would then be the product of speakers merely generalizing from their native language. Relatedly, the fact that all participants spoke English is also a limitation for our interpretation of the cross-linguistic corpus analysis. Because our complexity norms were elicited for English words from English speakers, the ratings are likely imperfect measures of conceptual complexity for words translated into other languages. Thus, it is difficult to know whether variability in the magnitude of the complexity bias cross-linguistically is due to true underlying differences in the bias, or merely a difference in the fidelity of the complexity ratings cross-linguistically. Speaking against this limitation, however, the presence of a complexity bias across all 80 languages that we examined suggests that the bias is likely to hold cross-linguistically in experimental work as well. If anything, the cross-linguistic mean bias is likely larger than our current estimates in the corpus study, because of the mismatch in complexity judgments between English speakers and speakers of other languages.

The motivating framework for the present work was the notion of interacting dynamics at multiple timescales. Our work suggests that a complexity bias is present in both individual speakers—the pragmatic timescale (Hypothesis 1)—and in the structure of the lexicon—the language evolution timescale (Hypothesis 2). While the existing data do not speak directly to a causal relationship between these two hypotheses, a casual interpretation is both parsimonious and consistent with work in other domains of linguistic structure, reviewed in the Introduction. A causal account would suggest that the trade off between listener and hearer pressures leads to a complexity bias at the pragmatic timescale and,

over time, these pressures lead to the same regularity emerging in the lexicon over the language change timescale. Our data are not able to speak to the processes underlying participants' judgments—these judgments need not reflect in-the-moment pragmatic inference; they could also be the result of an iconic mapping between effort and meaning, or a lower-level statistical regularity extracted through extensive experience with a language. Regardless of the cognitive instantiation of this inference, the result is lexicons that reflect Horn's principle.

Chapter 3

What is conceptual complexity?

(?, ?, ?, ?, ?, ?, ?)

OUTLINE DIFFERENT THEORIES. The goal is not to distinguish between these different theories, but rather to consider each as a framework for understanding conceptual complexity.

3.1 Experiment 1: Descriptions of objects

In Chapter 2, we presented participants with novel, real objects and measured their complexity through explicit judgements (Exp. 4; pg. 24) and study time (Exp. 8b; pg. 34). Both of these measures showed variability, suggesting that these objects differed in their complexity. However, these measures do not tell us *what* differs across objects; what makes one object more complex than another. The Classical Theory of concepts would suggest that these objects differ in complexity because they are composed of different number of conceptual primitives, with more complex objects containing more primitives that simpler objects.

Theory	Relevant Dimension	Prediction	Relevant Studies
Classical	# of primitives	Concepts with longer definitions (and thus more primitives) will be more complex.	Studies 1-3
Classical	Entropy of associates	Concepts with higher entropy of associates will be more complex.	Study 4
Exemplar	# of exemplars	Concepts with more exemplars will be more complex.	Studies 5-6
Prototype	# of exemplars	Concepts with fewer exemplars will have more uncertainty, and thus be more complex.	Studies 5-6
Prototype	Variability of exemplars	Concepts with more variable exemplars will be more complex.	Study 7

Table 3.1: Summary of studies.

In Experiment 1, we reasoned that, if true, these primitives should be reflected in participants linguistic descriptions of the objects. In particular, we predicted that objects rated as more complex and studied longer in Exp. 4 and 8b (Chapter 2) should also be described with longer descriptions. We tested this prediction by asking participants to produce written linguistic description of the objects.

3.1.1 Methods

Participants

In this and all subsequent experiments, participants were recruited on Amazon Mechanical Turk and received US $0.15-0.50 for their participation, depending on the length of the task. 60 participants completed this first experiment.

Stimuli

We used the same set of 60 novel real objects as in Chapter 2 (Fig. 2.3; pg. 25).

Procedure

On each trial, we presented a single object and following instructions: "Look at the object below. Imagine you just received this object as a gift. Describe what the object looks like to a friend." Participants then entered their description in a text box below the object.

Each participant described 10 objects in total. Five objects were from the top quantile (high complexity) and 5 objects were from the bottom quantile (low complexity). Order of objects was randomized.

3.1.2 Results and Discussion

Example descriptions for a sample low and high complexity object are presented in Table 3.2. We considered two measures of length: log number of words and log number of characters. Across objects, the mean length of description was $M = 8.82$ words ($SD = 1.14$) and $M = 36.31$ characters ($SD = 4.69$).

The key question was whether the description length was related to the psychological

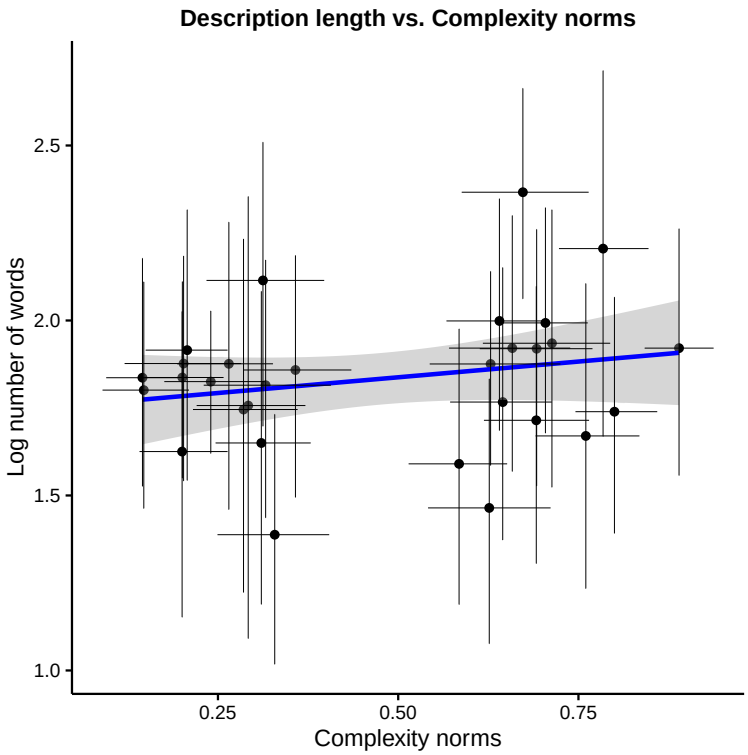

Figure 3.1: Relationship between description length and complexity norms. Error bars show 95% confidence intervals.

	Low-complexity object
	"cup holder" "it is a bowl with a black portion on top" "football kicker's stand" "a hole type thing" "it looks like an ash tray, but a bit shallow" "looks like a dog bowl"

	High-complexity object
	"a robotic carpet shampooer" "it's a flat silver disk on rollers with what appear to be tall handlebars standing away from it at an angle" "a pressurized floor buffer on wheels" "it looks like a high tech metal detector on wheel." "it kinda looks like a portable lamp" "It is a machine with a circular stand and wheels, it has a metal handle"

Table 3.2: Sample descriptions of a low- (top) and high- (bottom) complexity objects. Overall, descriptions were longer for high-complexity objects.

correlates of complexity measured in Chapter 2, explicit ratings and study times. To test this question we fit a linear mixed-effect model predicting log number of words with complexity norms as a fixed effect, and a second model predicting log number of words with study times as a fixed effect. As evident from Table 3.2, participants varied considerably in the syntactic construction of their descriptions as well as overall length, making it important to control for this variability using a mixed-effect model. The random effect structure included by-participant intercepts and by-trial slopes. There was a reliable relationship between log number of words and complexity norms ($\beta = .2$, $t = 2.8$; Fig 3.1), and between log number of words and log study time ($\beta = .59$, $t = 3.07$). The same pattern held for log number of characters.

In the context of the Classical Theory of concepts, this result suggests a connection

between conceptual complexity and number of primitives: More conceptually complex objects have longer descriptions, and thus more primitives.

3.2 Experiment 2a: Definitions of words

Experiment 1 suggested objects that appear visually more complex are described with longer descriptions. The studies in Chapter 2, however, suggest that the construct of conceptual complexity extends beyond visual complexity to abstract word meanings. This predicts the length of a dictionary definition of a word should be correlated with the conceptual complexity of its meaning. In light of the complexity bias observed in Chapter 2, we also predict that words with more complex definitions should be longer and have definitions that are rated as more complex. In Experiment 2 we tested these predictions by presenting participants with the definition of low frequency English words and asking them to rate the conceptual complexity of the definition.

We selected low frequency words for several reasons. First, because participants were unlikely to know the word associated with the definition, knowledge of a word's length was unlikely to affect the complexity judgement. Second, because the words were uniformly low frequency, this reduced the possibility that differences in word length were due to frequency, rather than conceptual complexity. Finally, because participants were unlikely to know the words, we could conduct a follow-up experiment (2b) probing judgements about the length of a meaning's word, without knowledge of the English word interfering with this judgement.

Word	Definition
bissextile	"a leap year"
mussitation	"movement of the lips as if in speech but without accompanying sound"
omphaloskepsis	"contemplation of one's navel as an aid to meditation"
parvis	"a court or enclosed space before a building"
sniddle	"long coarse grass"
zarf	"a holder, usually of ornamental metal, for a coffeecup without a handle"

Table 3.3: Sample definitions of real English words used in Experiment 2.

3.2.1 Methods

Participants

200 participants completed the task.

Stimuli

We selected 100 dictionary definitions of low-frequency words (see Table 3.3 for examples).

Procedure

The task was identical to Experiment 9 in Chapter 2 (pg. 34), except that participants were presented with definitions rather than words. Participants were first presented with instructions describing the norming task:

> In this experiment, you will be shown the definition of a word and asked to decide how complex the meaning is. A word's meaning is simple if it is easy to understand and has few parts. An example of a simple meaning is "brick." A word's meaning is complex if it is difficult to understand and has many parts.

An example of a more complex meaning is “engine.”

For each definition, we then asked “How complex is this definition?,” and participants indicated their response on a 7-pt Likert scale anchored at “simple” and “complex.” The first two words were always “ball” and “motherboard” to anchor participants on the scale. Each participant rated a sample of 10 definitions.

3.2.2 Results and Discussion

The central prediction is that definitions with more primitives in the definition, operationalized as the length of the definition, should be rated as conceptually more complex. To test this prediction, we fit a linear mixed-effect model predicting complexity ratings with log number of words in the definition as a fixed effect. The random effect structure included by-participant intercepts and by-trial slopes. As predicted, there was a strong relationship between complexity ratings and log number of words ($\beta = 1.50$, $t = 27.94$). The same pattern held for log number of characters.

A secondary prediction is that there should be a relationship between the length of the definition and the length of the word: If languages encode the conceptual complexity of a word’s meaning in the length of the word, longer words should be associated with longer definitions. This prediction was not supported ($r = .05$, $p = .59$). Finally, we should also expect that longer words should be associated with definitions rated as more complex. We fit the same mixed-effect model as above to test this prediction, except with log number of characters in the word as a fixed effect. There was a significant relationship between word length and rating judgements ($\beta = .52$, $t = 2.86$), suggesting more complex definitions area associated with longer words. However, in a model with both word length and definition length as fixed effects, word length was no longer a reliable predictor of complexity ratings

($\beta = .18, t = 1.19$).

In sum, we find a strong relationship between definition length and conceptual ratings, as predicted by Classical Theory of concepts: Longer definitions, with more primitives, are rated as more complex. We do not, however, find the predicted relationship between the conceptual complexity of the definition and word length, as would be predicted by the studies in Chapter 1. One possible explanation for this null finding is that participants' complexity ratings were driven by the linguistic complexity of the definition, rather than its conceptual complexity. In other words, participants may have rated longer definitions as more complex *because* they were longer, not because they were more conceptually complex. The current design does not allow us to distinguish these two possibilities.

3.3 Experiment 2b: Definition mapping

If definition length is related to conceptual complexity and participants have a complexity bias, we predict that participants should be biased to map more complex definitions on to longer words. In Experiment 2b, we test this prediction in an experiment analogous to the word mapping experiments in Chapter 2. Participants were presented with a meaning—a definition—and asked to guess the translation of the meaning in an alien language from two possible alternatives, one long and one short.

3.3.1 Methods

Participants

200 participants completed the experiment.

Stimuli

We used the normed definitions from Experiment 2a. The short novel words containing one syllable, and the long novel words contained three syllables. There were 10 short and 10 long novel words presented in random order.

Procedure

Participants were first presented with the following instructions:

> In this experiment, you will see the definition of a word. Your job is to guess what the translation of that word is in an alien language. You will make your guess by betting on two possible words in the alien language. Imagine you have a $100 dollars. To place your bet, assign an amount to each of the words. Your bets must add to 100.

Participants then viewed a definition and two possible alternative words, one short and one long. Participants selected a response by placing a numeric bet (0-100) under each word. Each participant rated 10 definitions in total.

3.3.2 Results and Discussion

Consistent with previous evidence, we found a complexity bias in participants mappings from definition to words: Participants tended to map definitions rated as more complex in Experiment 2a to longer words ($r = .39$, $p < .0001$). However, there was also a strong correlation between participants bets to longer words and definition length, measured in terms of log number of characters ($r = .82$, $p < .0001$). In an additive linear model predicting bets to the long word with both complexity norms and definition length, definition length

($\beta = 3.81$, $t = 2.64$, $p < .01$) was a significant predictor of bets, but complexity norms were not ($\beta = 1.02$, $t = 1.38$, $p = .17$).

While this result is consistent with a complexity bias, as well as the pattern of complexity predicted by the classical theory of concepts, it is difficult to make strong causal inferences from these data. The fact that definition length accounts for more variance in bets than complexity norms suggests that it may be linguistic complexity, rather than conceptual complexity that is driving this bias. This result may simply reflect participants bias to map long definitions to long words. This result, while a positive finding, does not speak to the claim directly that it is *conceptual* complexity per se that is related to definition length and the bias in word length. In Study 3, we try to address this issue more directly.

3.4 Study 3: Feature norms

The above approach is messy.

3.4.1 Methods

3.4.2 Results and Discussion

3.5 Study 4: Entropy of associates

3.5.1 Methods

3.5.2 Results and Discussion

3.6 Experiment 5a: Simultaneous frequency

We now turn to predictions about conceptual complexity from other theories of concepts: Exemplar and prototype theories. Both of these theories predict that aspects of experience should influence the representation of a concept, and thus the conceptual complexity of the concept. In particular, each makes predictions about the influence of observing individual exemplars of a concept. Importantly, however, these two theories make different predictions about the direction of this influence. Under the exemplar theory, a participant who observes more exemplars of a concept consequently would have a representation of the concept that contained more individual exemplars. This is the case of an expert: A person who is a bird expert, for example, would have many exemplars of birds as part of the concept bird. This might mean that this concept has overall higher fidelity, than for a non-bird expert. Thus, the exemplar theory predicts that concepts with more observed exemplars will be more complex.

In contrast, the prototype theory predicts that participants represent only summary

statistics over exemplars. If true, this means that the more exemplars a participant observes the less uncertainty there will be about the underlying concept; the concept prototype. Thus, the prototype theory predicts that the more exemplars a participant observes from a given concept, the *less* conceptually complex that concept will be. This prediction also falls out of an information theoretic account: Objects that appear more frequently are less surprising, and thus less conceptually complex.

In Experiment 5, we explore the possibility that the number of exemplars a participant observes influences the conceptual complexity of the concept. We present participants with either a few or many exemplars of a concept, and then ask participants to map that concept to either a short or long label. Under either theory, if conceptual complexity is related to exemplar frequency, we should expect the number of exemplars observed to influence how a concept is mapped to words of varying length. Under the exemplar theory, we expect long words to map to concepts with many exemplars, and under the prototype theory, we expect long words to map to concepts with few exemplars.

3.6.1 Methods

Participants

477 participants completed the experiment.

Stimuli

The objects were composed of a single geon, similar to those used in Expts. 1-3 in Chapter 2. The linguistic stimuli were novel words composed of either 2 (e.g., “tupa,” “gabu,” “fepo”) or 4 (e.g., “tupabugorn,” “gaburatum,” “fepolopus”) syllables.

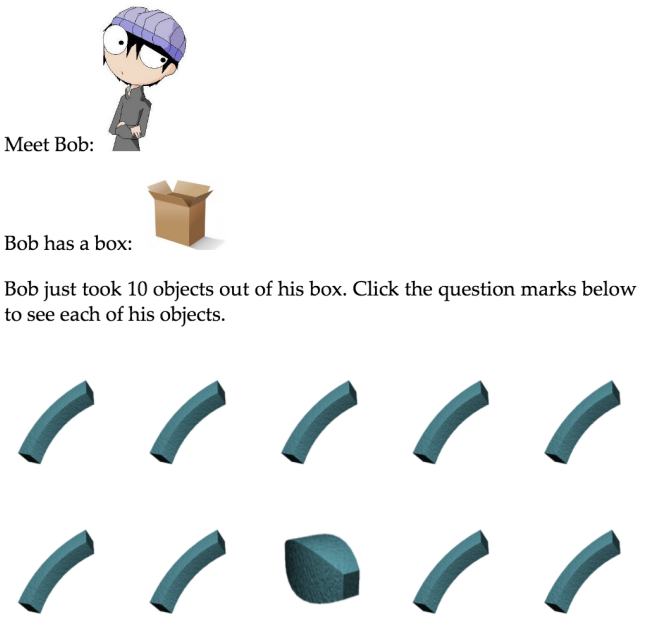

Figure 3.2: Sample display in training phase in Experiment 5a.

Procedure

We presented participants with 10 objects on a single screen with a cover story that the objects were from a character's box (see Fig. 3.2 for precise text). Participants were required to click on a question mark to reveal each object. There were always two types of objects, one appearing nine times and the other once. Order of presentation was randomized.

After this training phase, participants completed a forced choice mapping task, as in Studies 1 and 5 in Chapter 2 (pg. 17 and pg. 25). We presented a short or long word and asked participants to make a judgment about whether the word referred to the low or high frequency object. Each participant completed a single mapping trial, and word length was manipulated between participants.

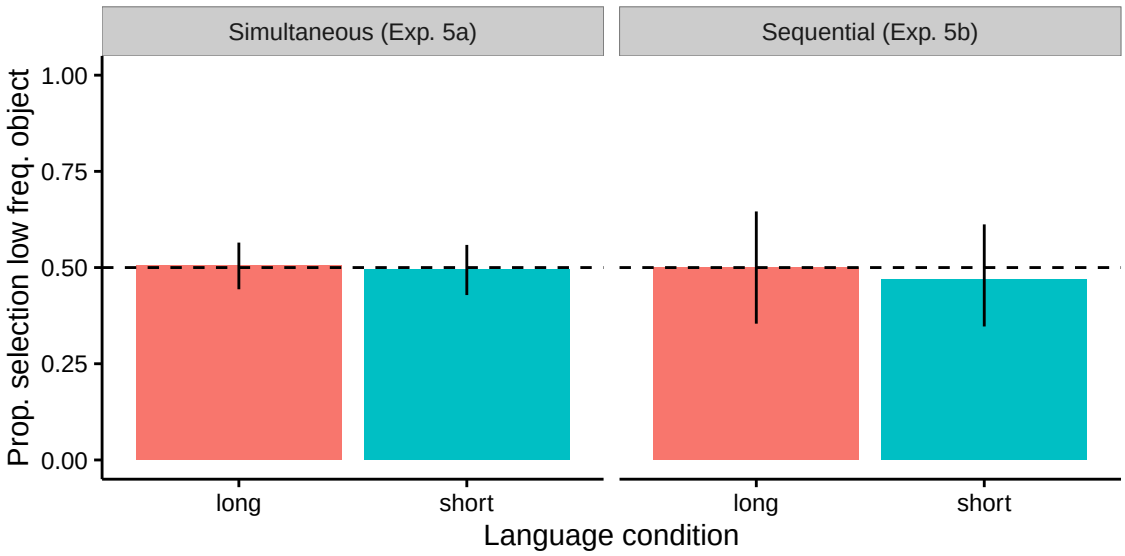

Figure 3.3: Proportion participants selecting the low frequency object as a function of language condition, in Exp. 5a (left) and Exp. 5b (right). Error bars are bootstrapped 95% confidence intervals.

3.6.2 Results

Selections between the two conditions did not differ ($\chi^2(1) = 0.02$, $p = .89$; Fig. 3.3, left).

3.7 Experiment 5b: Sequential frequency

Given the null finding in Experiment 5a, we next explore whether the timing of the presentation of exemplars affects complexity. Previous work has shown that participants may induce different concepts depending on whether or not exemplars are presented simultaneously, as in Exp. 5a, or sequentially (e.g. ?). In Experiment 5b, we present exemplars sequentially such that only one exemplar is visible at any given time.

3.7.1 Methods

Participants

97 participants completed the experiment.

Stimuli

The objects were the set of novel, real objects used from Chapter 2, Expts. 4-6.

Procedure

We manipulated object frequency by sequentially presenting objects. Participants saw 60 objects one at a time for 750 ms per object. One object was presented 10 times and a second object was presented 40 times. Ten additional objects were included as fillers, each appearing once. These were included to make the critical manipulation less obvious. Order of presentation was randomized. After this training phase, participants completed a single mapping trial as in Experiment 5a. Word length was manipulated between participants.

3.7.2 Results and Discussion

Selections between the two conditions did not differ ($\chi^2(1) = 0.01$, $p = .92$; Fig. 3.3, right). Across Expts. 5a and 5b, we find no evidence that the frequency of exemplars is related to the conceptual complexity of objects.

3.8 Experiment 6: Facts

In Experiment 6, we again test the hypothesis that the number of observed exemplars is related to conceptual complexity, as predicted by both the exemplar and prototype theories.

Category	Fact
facebook	"_____s appear in the background of {many/a couple} pictures on Facebook."
money	"_____s cost {$4/ $400}."
talk	"_____s get talked about once a {day/year}."
use	"_____s are used once a {day/year}."

Table 3.4: The eight facts used in Experiment 6. For each of the four categories, there was a high and low frequency alternate (presented in curly brackets).

However, in the present experiment, we explore the possibility that perhaps raw number of times an exemplar is observed is not the psychologically relevant dimension of frequency. We instead ask whether *conceptual* frequency, or markedness, is related to conceptual complexity by teaching participants facts about novel objects.

3.8.1 Methods

Participants

120 participants completed this experiment.

Stimuli

All participants were presented with the same eight facts, shown in Table 3.4. Objects were again a sample of the novel, real objects used in Chapter 2, Expts. 4-6.

Procedure

Participants were presented with the following instructions:

> In this experiment, you will learn facts about objects but the name of the object will be missing. For example, a fact like, "Mops are used for cleaning the

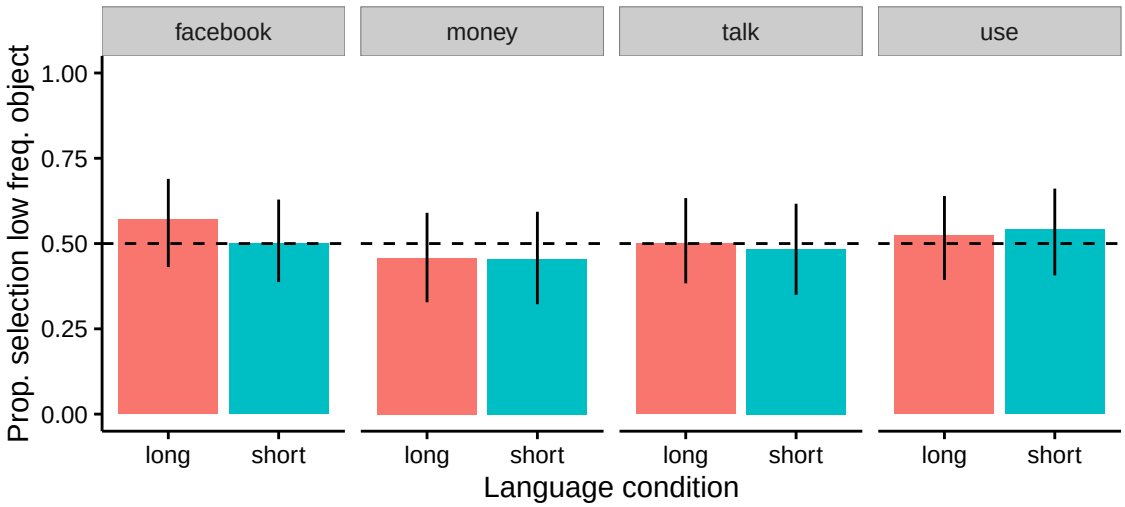

Figure 3.4: Proportion participants selecting the object associated with the low frequency fact as a function of fact category and language condition. See Table 3.4 for explanation of fact categories. Error bars are bootstrapped 95% confidence intervals.

floor," will appear as "_____s are used for cleaning the floor."After you memorize these facts, you will then be asked to guess the name for each object.

Learn 8 facts, repeat until you know it Four mapping trials everything randomized.

3.8.2 Results and Discussion

Across all four categories, there was no evidence that selections to the short or long word varied as a function of the frequency valence of the fact (all $\chi^2(1) < 3.8$; $p > .5$). In sum, then, across both Experiment 5 and 6, we find no evidence that exemplar frequency, either objective or psychological, is related to conceptual complexity. Next we turn to a final hypothesis about the nature of conceptual complexity: exemplar variability.

3.9 Experiment 7: Exemplar variability

Finally, we test an alternative hypothesis about conceptual complexity: A concept is conceptually complex if its exemplars are highly variable. This is predicted by

cite haskell

3.9.1 Methods

Participants

Stimuli

Procedure

3.9.2 Results and Discussion

Chapter 4

Origins of a complexity bias

4.1 Introduction

A universal property of languages is that they contain units of meaningful sounds—words—that vary in length. What accounts for this variability? That is, why is the word for "can" short but the word for "calculator" long? One class of explanations for this variability appeals to properties of the linguistic form itself, such as word frequency (Zipf, 1936) and predictability in linguistic context (Piantadosi et al., 2011a; Mahowald, Fedorenko, Piantadosi, & Gibson, 2013). Our recent work has revealed an additional factor influencing word length: conceptual complexity. Across 80 natural languages, we find a bias for longer words to refer to conceptually more complex meanings (a *complexity bias*; Lewis, Sugarman, & Frank, 2014). This systematicity between word length and meaning challenges the long-held assumption that the relationship between form and meaning is entirely arbitrary (Saussure, 1916, 1960).

The origins of this bias in language are an open question. One possibility is that the bias is due to a pressure in individuals to map longer words onto more complex meanings.

	Object								
	1 (Q1)	2 (Q1)	3 (Q2)	4 (Q2)	5 (Q3)	6 (Q3)	7 (Q4)	8 (Q4)	9(Q5)
Gen. 0	damitobup	nagir	nid	gimunobugup	dunobax	mikupudax	bipag	daganitobip	nimimog
Gen. 1	nilobup	niger	nid	runtunbug	dunobug	bipoxtog	bipag	dipentag	nimimog
Gen. 2	nilobup	niger	nid	runtunbug	dunbug	ripenbog	bippenbog	dipentag	nimobop
Gen. 3	nilobop	niger	nid	rittenbob	dabop	rudentag	buttenbug	dertag	nimobop
Gen. 4	nilobop	niger	nid	bittenbob	dabop	rittenbog	buttenbop	dertag	nimbobop
Gen. 5	nilop	niger	nir	girbop	dabop	dirbop	bittenbop	rittenbog	nilobop
Gen. 6	nilop	niger	nir	garbog	dabop	dabog	bittenbop	rittenbog	nilop
Gen. 7	nilop	niger	hir	garbop	dabog	dabog	bittenbop	rottenbog	nilop

Table 4.1: A representative language chain. Words are presented for each of the 10 objects across 7 generations and the initial input language. The complexity quintile of the object is noted parenthetically. Across generations, words tend to get shorter, less unique, and phonotactically more probable. Words also become more likely to be remembered accurately.

Under this account, there is a psychological bias to map longer words onto more complex meanings—a synchronic complexity bias—and over time this bias leads to this same regularity emerging in the structure of the lexicon—a diachronic complexity bias. In the present paper, we consider the mechanism through which a synchronic complexity bias in individuals might lead to diachronic change in the lexicon.

There are several possible sources for a psychological, synchronic complexity bias. For example, the bias could reflect a more general cognitive preference for iconicity (see Schmidtke et al., 2014, for review). A second alternative is that the bias is related to principles of communication. As part of a broader theory of communication, Horn (1984) suggested that a contrast in length between two phrases with the same denotational value implies a contrast in meaning, with the longer phrase getting the more unusual or complex meaning. Thus, the complexity bias in the lexicon could reflect this in-the-moment communicative bias—an appealing possibility given evidence that other features of the lexicon also reflect principles of communication, like the structure of semantic space (Regier et al., 2007; Kemp & Regier, 2012; Piantadosi, Tily, & Gibson, 2012).

Critically, if the emergent diachronic bias is due to a psychological synchronic pressure, we should be able to observe this bias not only in the structure of natural languages, but also in one-shot learning tasks with novel words. In previous work, we have found robust support for this prediction. Across a range of stimuli, and both comprehension and production tasks, we find that speakers are biased to map a longer novel word onto a more complex novel referent, relative to a shorter word (Lewis et al., 2014).

How does a synchronic complexity bias lead to diachronic change in the lexicon? The causal mechanism for this type of change would have to take place over multiple timescales: A synchronic bias in the moment of language interaction would have lead to changes in the lexicon over the course of language evolution. We propose that a psychological bias causes small changes in memory for complex phonological forms in the moment of language interaction, and this pressure leads to biases in linguistic transmission across generations. Over the course of language evolution, these psychological, synchronic biases result in a lexicon that magnifies these biases (Griffiths & Kalish, 2007).

In the present work, we begin to test this hypothesis using the iterated learning paradigm, a recently-developed method for studying language change in the lab (e.g., Kirby, Cornish, & Smith, 2008; Reali & Griffiths, 2009; Smith & Wonnacott, 2010). The critical feature of this paradigm is that the learning output of one speaker becomes the learning input for a new speaker. This paradigm allows us to examine the evolution of a language for a "chain" of speakers learning and transmitting a language. The dynamics of these chains serve as an approximation of the dynamics of generations of children acquiring and then transmitting language to future generations.

A secondary goal of the present work is to examine how psychological pressures influence the structure of the lexicon, independent of conceptual pressures. Forms that are

difficult to remember are unlikely to survive in the language (Christiansen & Chater, 2008), and there may be an additional communicative pressure for economy of expression (Zipf, 1949). Both of these pressures might lead to a preference for shorter words over longer, harder-to-produce words, biasing the ultimate structure of the lexicon towards shorter, more memorable words.

We used an iterated learning paradigm to study the dynamics of these two aspects of the lexicon: how words change over the course of language evolution and how conceptual complexity interacts with these changes.[1] As predicted, we find that forms in the lexicon converge to a more stable state and that a complexity bias emerges in the mappings between words and referents. We also find, contra our hypothesis, that the complexity bias is attenuated over time. A post-hoc analysis suggests that this change in the complexity bias over time is related to the degree of cross-generational change in the lexicon.

4.2 Experiment

Given existing evidence that a complexity bias is present in one-shot learning games (Lewis et al., 2014), our experiment was designed to test how conceptual pressures influenced the lexicon over the course of transmission. We asked speakers to learn a novel language that contained meanings of varying complexity and words of varying length. Critically, the language we asked participants to learn contained no systematic relationship between complexity and word length. After studying these mappings, participants were asked to recall them. The measure of interest was the relationship between the errors participants made and the complexity of the referent. If participants show a complexity bias, they should

[1] For ease of measurement, we operationalize word length in terms of number of orthographic characters. However, this measure is highly correlated with measures of length with greater psychological reality, such as phonemes and morphemes (Lewis et al., 2014).

be more likely to add characters for more complex objects and remove characters for less complex objects.

This design characterized the first generation of our task. We then gave the labels that participants produced in the test phase of this first generation to a new set of speakers and asked them to complete the exact same task. We iterated 7 generations of this task in total.

4.2.1 Method

Participants

We recruited 350 participants from Amazon Mechanical Turk. Each generation was composed of 50 learners.

Stimuli

The referents were a set of 60 real objects that did not have common labels associated with them. These objects had been normed for their complexity in previous work (Lewis et al., 2014, Figure 1). Norms were obtained by asking participants to indicate "How complicated is this object?" using a slider scale. Norms were highly reliable across two samples of 60 participants. Based on these norms, we divided the objects into quintiles of 12 objects each. Each participant saw 2 objects from each quintile.

In the first generation, the words were composed of randomly concatenated syllables of 3, 5, 7, 9 or 11 characters in length. Words contained CV syllables and ended in a consonant (e.g., "gan," "panur," "pugimog," "tigadogog," and "mogonokigan"). Each participant saw 2 words of each length. The assignment of word lengths to objects was arbitrary.

Participants in Generation 2 were yoked with a participant from this first generation. This meant a participant in Generation 2 would see the exact same set of pictures as the

yoked participant from Generation 1, but would learn the labels for those objects that the yoked participant had produced in the testing phase of Generation 1. Order of presentation in the training phase was randomized across generations. We iterated this procedure for a total of 7 generations.

Procedure

Participants viewed a webpage that informed them they would be learning the names of 10 objects in an alien language. They were told they would see the names for each object four times and then their memory for the name of each object would be tested. Participants next viewed a screen displaying an object and the associated label below it. Participants pressed the space bar to advance to the next picture. Each picture-word pair was shown four times.

In the test phase, participants saw a screen with a picture and were asked to type the learned label in a text box below the picture. Memory for each of the 10 objects was tested.

Figure 4.1: Object stimuli used in the Experiment. The objects are sorted from least complex (top left) to most complex (bottom right) based on the complexity norms in Lewis et al. (2014). Each row corresponds to a quintile.

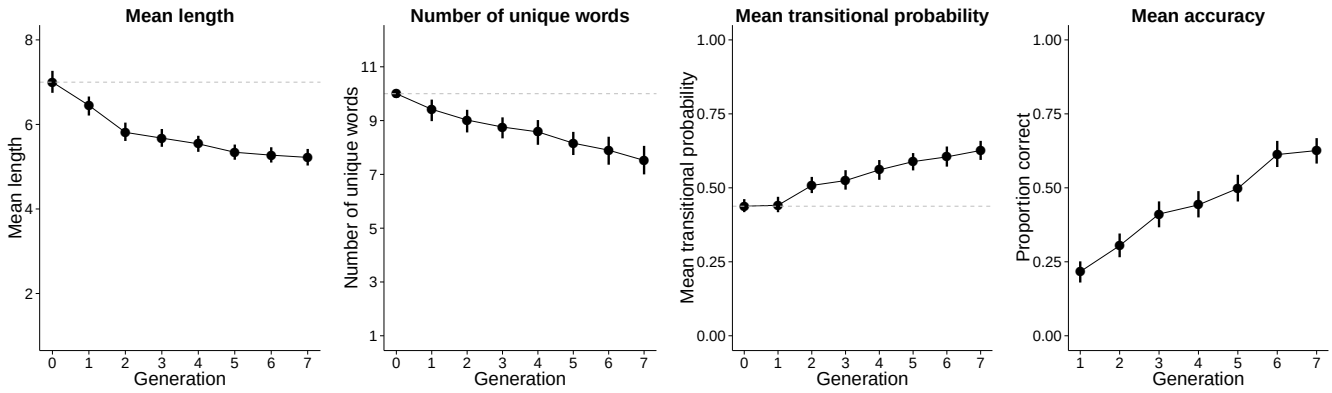

Figure 4.2: Changes in lexical features across generations. Error bars represent 95% confidence intervals computed via non-parametric bootstrap across chains.

4.2.2 Results

We conducted three analyses exploring how iterated learning influenced the structure of lexicons.[2] In Analysis #1, we examined the evolution of lexical forms. In Analysis #2, we considered the relationship between word length and referent complexity. This was the key analysis because it allowed us to test for a complexity bias in the lexicon and how this bias changed over time. Finally, in Analysis #3, we conducted a post-hoc analysis to understand the source of variability in cross-generational change in complexity bias across chains.

Across generations, 1% of object labels were excluded because they contained more than one word or no word was produced. In these cases, the object was re-assigned a label from a different participant in that generation. The label was selected from a trial that had both the same initial word length and an object from the same quintile.

[2] All code and data for the paper are available at `http://github.com/mllewis/iteratedRC`.

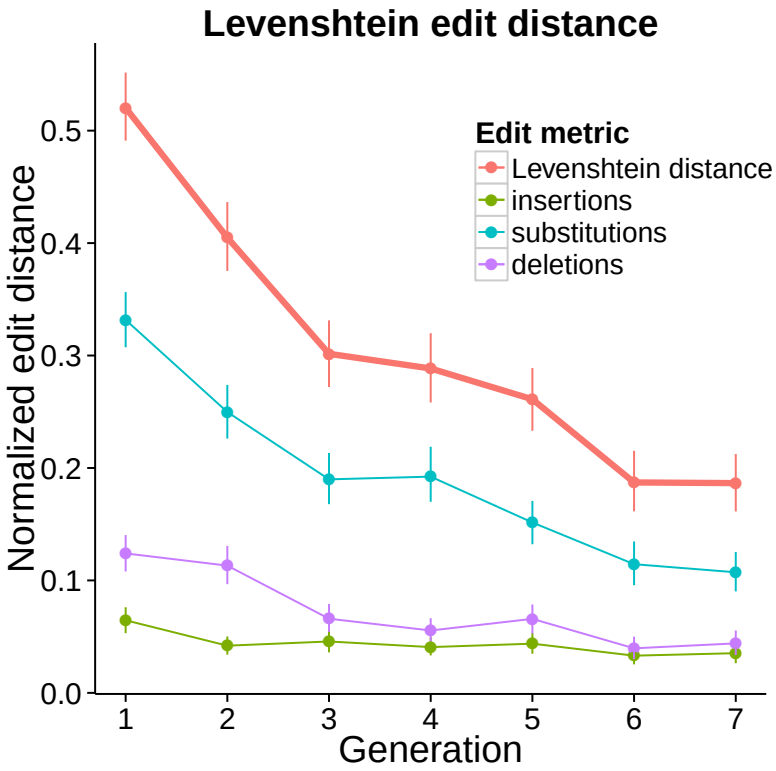

Figure 4.3: Edit distance across generations, normalized by length of the longest word (guessed word vs. actual word). The top line shows the Levenshtein edit distance. The lines below reflect the components of this metric (substitutions, deletions, and insertions). Error bars represent 95% confidence intervals computed via non-parametric bootstrap across chains. Number of edits decreased across generations.

Analysis #1: Word forms

Table 4.1 presents a representative language chain. We analyzed four features of the lexical forms, averaging across each of the 50 chains at each generation: mean word length, number of unique words, transition probability, and accuracy. We also analyzed the degree of lexical change at each generation using the Levenshtein edit distance metric.

Across generations, mean word length decreased from an initial length of 7 characters to 5.22 characters in Generation 7 ($SD = 2.25$; $r = -0.22, p < .0001$; Figure 4.2a). The

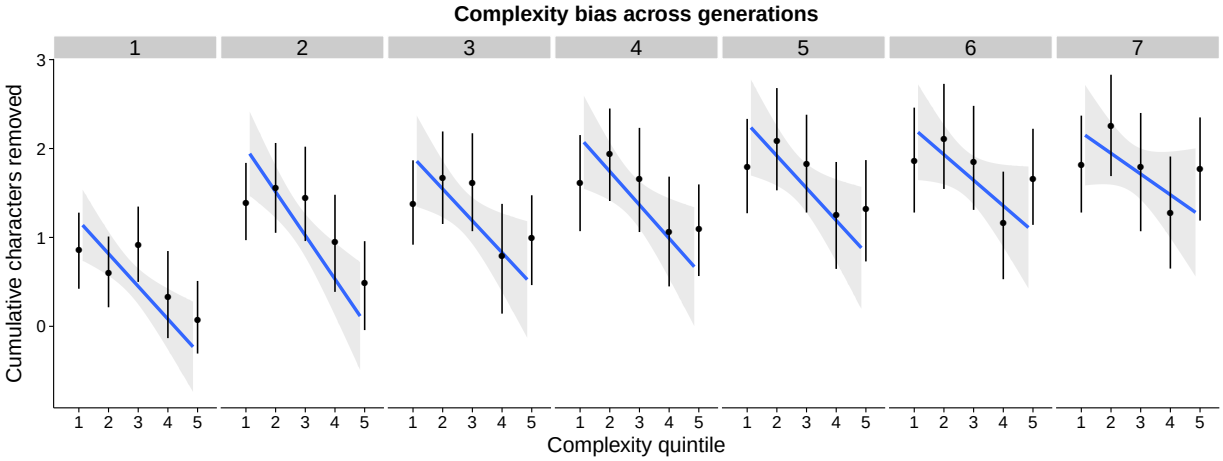

Figure 4.4: Cumulative characters removed as a function of complexity across all 7 generations. Points correspond to the quintile means. Lines represent the best fitting linear model predicting word length from the complexity norm of the object. Negative slopes indicate a bias to recall longer labels for more complex objects. Across generations, this bias decreased.

number of unique words also decreased across generations ($r = -0.35, p < .0001$; Figure 4.2b). Lexicons tended to reduce in size by mapping the same word to multiple objects (e.g., in the chain presented in Table 4.1, "nilop" refers to both Objects 1 and 9).

Third, the mean orthographic transition probability of each word increased across generations ($r = .52, p < .0001$; Figure 4.2c). Transition probabilities were calculated based on the set of words in the lexicon for a particular participant at a particular generation. This finding suggests that lexicons became more phonotactically structured across time. We also calculated the mean transition probability of each word using English transitions. Probabilities were estimated via orthographic bigrams from the Google Books corpus (Norvig, 2013). In this analysis, the mean English transition probability of each word also increased across generations ($r = 0.18$, $p < .001$), suggesting that the orthographic structure of individual words became somewhat more similar to English across generations.

Fourth, we found that participants became more accurate in recall across generations

($r = .46, p < .0001$; Figure 4.2d). To examine the relationship between accuracy and word forms, we constructed a logistic mixed-effects model predicting accuracy with word length, word uniqueness, and transition probability.[3] Only word length was a reliable predictor of accuracy ($\beta = 1.21,\ p < .0001$), suggesting that perhaps the increase in accuracy across generations was due to the shorter length of the words in these languages.

Finally, we analyzed word changes across generations using Levenshtein edit distance. This measure provides a formal metric of the similarity between two strings. Levenshtein edit distance is computed by counting the minimum number of character edits necessary to transform one word into another. For example, the edit distance from "can" to "cat" is 1 (1 substitution), while the edit distance from "can" to "calculator" is 8 (1 substitution and 7 insertions). For each word, we calculated a normalized measure by dividing the edit distance between the guessed word and the actual word by the length of the longest of the two. This normalized measure controlled for the decrease in word length across generations. Across generations, the normalized edit distance decreased ($r = -.30, p < .0001$; Figure 4.3). This decreasing trend also held for each of the components of the Levenshtein metric: number of deletions ($r = -.18, p < .0001$), insertions ($r = -.08, p < .0001$) and substitutions ($r = -.27, p < .0001$).

Taken together, this set of analyses points to a lexicon that is evolving to become more regular and consequently easier to learn.

Analysis #2: Complexity bias

In Analysis #2, we examined the relationship between changes in word length and the complexity of referents. If there is a complexity bias in the lexicon, participants should be

[3]The model specification was as follows: `accuracy ~ guessed label length × transition probability × uniqueness + (guessed label length | subject) + (1 | chain)`.

Quintile #2	Quintile #1: 2	3	4	5
1	86	78	64	52
2		84	63	34
3			41	59
4				58

Table 4.2: Contingency table of trials where participants recalled the same word for multiple objects. Columns correspond to the complexity quintile of the target object and rows correspond to the complexity quintile of the object with the same word. The diagonal is excluded because the experimental design restricted the number of possible confusions for these cases (1 possible alternative vs. 2 for all other quintiles). In cases of confusions, participants tended to reuse a word from an object in a nearby quintile.

more likely to produce longer labels for more complex referents.

We considered two metrics of word length: Label length in characters and cumulative characters removed (CCR). CCR is calculated by subtracting the word length at a particular generation from the input generation word length. Though slightly more complex, CCR provides a length metric that controls for variability in input word length; this control is important because words varied dramatically in their initial length due to random assignment in the initial generation. We calculated p-values based on an empirical distribution of r-values, obtained by sampling from random pairings of words and objects. This was done because changes in language forms across generations change the distribution of possible r-values.

Across generations, there was a reliable bias to map longer words to more complex referents across both measures of length (label length: $r = .05, p < .05$; CCR: $r = -.11, p < .0001$). Figure 5.2 shows CCR as a function of object complexity across generations. Qualitatively, the bias decreased across generations. However, there was high variability across chains both in the total complexity bias (label length: $SD = .27$), and in how this bias changed across generations (label length: $M = .004$; $SD = .69$).

A number of other exploratory analyses suggest a role for complexity in language change. First, Levenshtein edit distance was systematically related to the complexity of referents: Participants were more likely to edit words referring to more complex referents ($r = .05, p < .01$). Second, there was systematicity in the kinds of errors participants made when reusing words across multiple objects. Participants tended to reuse labels from objects of nearby quintiles (Table 4.2), suggesting that these labels were more conceptually confusable and lead to more category-formation.

Together, this set of analyses replicates prior work suggesting a complexity bias in the lexicon: Across both measures of word length, participants tended to recall longer labels to refer to more complex referents. They were also more likely to edit words related to more complex referents and reuse labels of objects from nearby quintiles. However, an unexpected finding was the attenuation of this bias across generations. In our last analysis, we try to understand this trend.

Analysis #3: Relationship between change in word forms and change in complexity bias

We conducted a post-hoc exploration of the variability in the complexity bias across chains. For each chain, we quantified the complexity bias at each generation by calculating the correlation between metrics of length (label length and CCR) and the complexity norms. We then calculated the correlation between these coefficients and generation. This gave us a measure of the change in the complexity bias across generations. We considered how this change in complexity bias related to the degree of change in the forms of the lexicon. Two metrics of lexical change were analyzed: accuracy and Levenshtein edit distance.

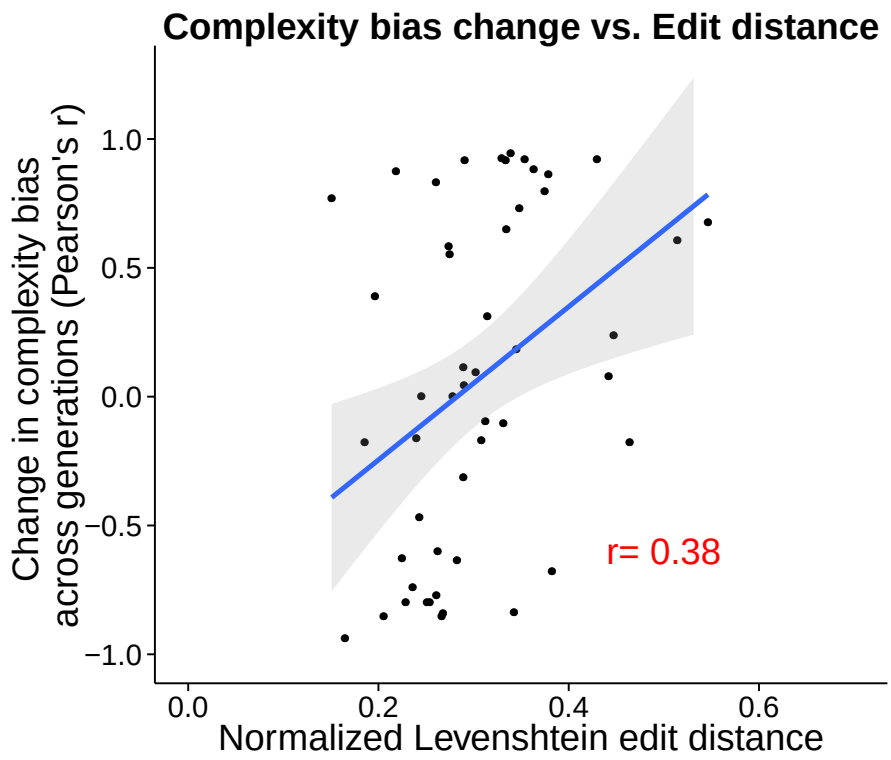

Figure 4.5: Complexity bias as a function of the normalized Levenshtein edit distance of the chain. Complexity bias is calculated here using number of cumulative characters removed. Each point corresponds to an individual chain. Chains with greater normalized Levenshtein distances tended to show a greater increase in complexity bias across generations.

Chains with greater cross-generational change in lexical forms tended to show an increase in complexity bias over time. Using raw label length as the length metric, there was a reliable correlation between change in complexity bias and accuracy ($r = 0.29$, $p < .05$) and between change in complexity bias and normalized Levenshtein edit distance ($r = -0.32$, $p = .02$). This same pattern also held for the CCR length metric (accuracy: $r = -0.37$, $p < .01$; Levenshtein: $r = 0.38$, $p < .01$; Figure 4.5).

4.2.3 Discussion

In three analyses, we examined change in the structure of lexicons across generations of transmission. Analysis #1 reveals that lexical forms become simpler and more regular over time. We find that words become shorter, less unique, more phonotactically probable, and more likely to be remembered. We also find that this structure facilities memory recall: lexicons with fewer and shorter words are more likely to be remembered accurately. Analysis #2 examined the relationship between lexical forms and conceptual structure, and found that a complexity bias emerges in the lexicons.

An unpredicted result was that the complexity bias does not strengthen across generations. Analysis #3 suggests that change in the complexity bias across generations is related to the degree of change in lexical forms in the chain: Chains with more change are more likely to show an increase in complexity bias over time. The underlying mechanism supporting this relationship is straight-forward: chains that make more errors have more opportunity to deviate from the random input mappings between words and referents. This direction of this correlation suggests that when chains do in fact deviate from these initials mappings, they do so in a systematic way. That is, they tend to deviate in a way that is more likely to map longer words onto more complex referents.

4.3 General Discussion

The iterated learning paradigm provides an opportunity to examine how in-the-moment psychological pressures influence the structure of a language in aggregate, over time. We examined two aspects of this structure: lexical forms and the mappings between words and objects. We hypothesized that different psychological pressures would influence each type

of structure. In the case of lexical forms, we predicted there would be a bias to simplify the language into shorter, fewer forms. In the case word-object mappings, we predicted a bias to map longer words onto more complex meanings (Lewis et al., 2014). The question of interest was how these psychological pressures influenced the structure of the lexicon across generations of transmission.

Our findings suggest that each of these pressures may have influenced the structure of the lexicon—and critically—that they interacted with each other. We found both a bias to simplify the lexicon and a bias to map longer words onto more complex meanings. But these pressures appear to have pushed in opposite directions: The pressure to simplify the language leads to less variability in word length, and this reduced variability suppresses the complexity bias.

If these dynamics reflect actual language evolution, however, an important question still remains—why do we in fact see a complexity bias in natural language? That is, if there is a strong pressure towards simplicity, then why does a complexity bias emerge in natural language despite this pressure?

One possibility is that this discrepancy is due to the absence of an important feature in our task: communication with a second interlocutor. Zipf (1949) argued that the equilibrium that emerges in the lexicon is a product of both the speaker's desire to say less and the listener's desire for a more explicit, comprehensible message. Importantly, the common desire for efficiency creates opposing pressures among interlocutors. For a speaker, the optimal solution to communication is to have a lexicon that contains a single, short word that can be used to refer to all meanings. However, for a listener, the optimal solution is to have a lexicon that maps a unique word onto every possible meaning.

Thus, perhaps the absence of a listener pressure in our task may have lead our participants ("speakers") to simplify the language. While our task was posed as a memory task, there was no penalty for failure to remember a form. In contrast, in a communicative task, the listener's failure to comprehend a label would have acted as an incentive for accurate reproduction, perhaps limiting the amount of compression the language would undergo.

But we speculate that memory limitations also play another role in the evolution of the lexicon: by introducing variation into the representations of individual words, speakers' memory constraints allow for change. In the absence of memory constraints, speakers might simply reproduce the language as is; thus, the interaction between cognitive and communicative pressures may function to *facilitate* the emergence of a complexity bias. This synergistic relationship between memory and change is reminiscent of the "less-is-more" hypothesis and its descendants (Newport, 1990; Hudson Kam & Newport, 2005), in which cognitive limitations are invoked as an important mechanism in language learning and language change. In the case of the complexity bias, these proposals make testable predictions that can be explored by extending the present paradigm into a communicative domain with varying demands on memory.

Chapter 5

Pressures shaping the evolution of the lexicon

5.1 Introduction

What factors shape language? Psychologists have made significant progress understanding this question in the domains of communicative interaction and children's developmental trajectories. In both cases, accounts rely on positing two pressures on the cognitive system—one internal and one external. In the case of communication, theorists argue that speakers are influenced by cognitive constraints (minimize effort) and by the needs of the communicative partner (be understandable; Horn, 1984). In the case of acquisition, there are internal maturational constraints, as well as external pressures from the quality and quantity of linguistic input (Hart & Risley, 1995). In the present paper, we explore the possibility that the same two pressures—system internal and external—may also shape *language systems* over the course of language evolution.

Central to this hypothesis is the notion of a timescale: there are different units of time

over which processes operate, and processes at shorter timescales influence those at longer timescales (Blythe, 2015, see also Fig. 1). At the shortest timescale are individual utterances in communicative interactions (pragmatics). At a longer timescale is language acquisition. Both experimental and modeling work suggest that communicative interactions at the pragmatic timescale influence processes like word learning at the acquisition timescale (e.g., Baldwin, 1991; McMurray et al., 2012; Frank, Goodman, & Tenenbaum, 2009; Frank & Goodman, 2014).

In addition to pragmatics and acquisition, a third relevant timescale is language evolution: the timescale over which entire language systems change. As for acquisition, there is evidence that language systems may be the product of processes at the pragmatic timescale. For example, languages universally structure semantic space to reflect optimal equilibria between communicative pressures (e.g., Kemp & Regier, 2012; Regier et al., 2007; Baddeley & Attewell, 2009).

However, the presence of communicative pressures at the pragmatic timescale is unable to explain cross-linguistic variability in linguistic structure. That is, why does Polish have rich morphology but English relatively sparse? A growing body of work argues that this

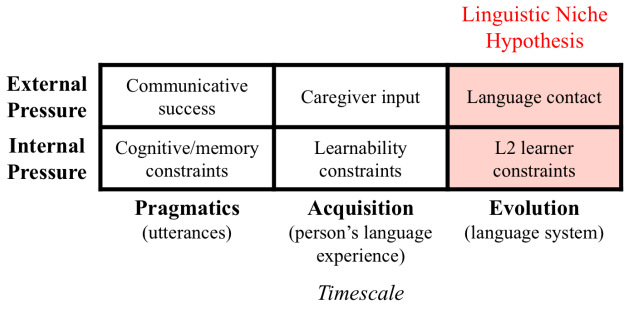

Figure 5.1: Pressures on language, internal and external to the cognitive system at three different timescales. The Linguistic Niche Hypothesis suggests that language evolution is influenced by the internal and external pressures in the particular environmental context in which a language is spoken.

variability may be due to cognitive constraints internal to the language learner (Chater & Christiansen, 2010) as well as properties of the environmental context (Nettle, 2012). This hypothesis, termed the *Linguistic Niche Hypothesis* (Lupyan & Dale, 2010; Wray & Grace, 2007), suggests that language systems adapt to the internal and external pressures of the linguistic environment.

A number of recent studies provide correlational support for this proposal. At the lowest level of the linguistic hierarchy, languages with larger populations are claimed to have larger phonemic inventories (Atkinson, 2011; Hay & Bauer, 2007), but shorter words (Wichmann, Rama, & Holman, 2011). Speakers with more second language learners have also been suggested to have fewer lexical items (Bentz, Verkerk, Kiela, Hill, & Buttery, 2015). At the level of morphology, speakers with larger populations tend to have simpler morphology (Lupyan & Dale, 2010; Bentz & Winter, 2013). Finally, there is also evidence that population size may influence the mappings between form and meaning. In particular, this work suggests that languages tend to map longer words to more complex meanings (Lewis et al., 2014), but that this bias is smaller for languages with larger populations (Lewis & Frank, 2016).

The plausibility of the Linguistic Niche Hypothesis depends largely on the presence of a possible mechanism linking environmental features to aspects of language systems. A range of proposals have been suggested (Nettle, 2012). For example, one possibility is that children (L1) and adult (L2) language-learners differ in their learning constraints. In particular, children may be better at acquiring complex morphology than adults, and so languages with mostly children learners may tend to have more complex morphology. A second possibility is that speakers in less dense social networks have less variable linguistic input, and this leads the language system to have more complex morphology.

Providing evidence for these mechanisms is empirically challenging, however. Because there are many factors that shape a linguistic system, large datasets are needed to detect a correlation with environmental factors. In addition, there is non-independence across languages due to genetic relationships and language contact, and so data from a wide range of languages are needed to control for these moderators (Jaeger et al., 2011). Third, the hypothesized mechanisms linking languages to their environments are somewhat underspecified. Finally, the large scale of this hypothesis makes it difficult to directly intervene on, and so we must rely primarily on correlational data to make inferences about mechanism.

In this work, we try to address some of these challenges by clarifying the empirical landscape. We do this by aggregating across datasets that find covariation between environmental variables and linguistic structure. This serves two purposes. First, it allows us to examine the relationship between the same set of environmental predictors across a range of linguistic features. And, second, it allows for the same analytical techniques and areal controls to be used across datasets. By addressing these inconsistencies, we are better able to compare directly relationships between environmental and linguistic features. A more coherent picture of the empirical landscape may in turn provide insight into the mechanism linking language systems to their environments.

We also explore a novel aspect of the Linguistic Niche Hypothesis: the relationship between L1 and L2 learnability. We ask whether the same languages that are more easily learnable by second language learners are also easier to learn for first language learners, or whether there is some tradeoff in learnability. As a proxy a language's learnability for child-learners, we use the mean age of acquisition of children's first words in a language.

In what follows, we first present a study examining the relationship between environmental and linguistic features using the same analytical techniques across all variables (Study 1). In Study 2, we examine the relationship between first language learnability and environmental and linguistic features.

5.2 Study 1: Environmental pressures on language

The Linguistic Niche Hypothesis suggests that languages are shaped by their environment, but the exact nature of these effects has varied across the literature—both in terms of the variables considered and the direction of the effect. To explore this variation, we combined data from five existing datasets that included environmental or linguistic data. The datasets were selected for being publicly available and containing a large sample of languages. Below we describe each of these datasets, followed by our analytical methods, and results.

5.2.1 Datasets

Lupyan and Dale (2010). This dataset contains grammatical information from WALS (Dryer & Haspelmath, 2013), and demographic and geographic information from Ethnologue and the Global Mapping Institute (Gordon, 2005). The demographic and geographic variables included total population of speakers, number of neighboring languages, area of region in which the language is spoken (km^2), mean and standard deviation temperature ($celsius$), and mean and standard deviation precipitation (cm). We used these data to create a metric of morphosyntactic complexity calculated from 27 of the 28 morphosyntactic variables analyzed in the original paper.[1] For each variable, we coded the strategy as simple

[1] WALS variable 59 was missing from the dataset.

if it relied on a lexical strategy or few grammatical distinctions (e.g., 0-3 noun cases), and complex if it relied on a morphological strategy or many grammatical distinctions (e.g., more than 3 noun cases). We summed the number of complex strategies to derive a measure of morphosyntactic complexity for each language, including only languages with data for all 27 variables[2]. [n = 1991 languages]

Bentz et al. (2015). Two variables were used from this dataset: ratio of L2 to L1 speakers and number of word forms. Estimates of number of word forms were taken from translations of the *Universal Declaration of Human Rights*. Number of word forms was calculated as the number of unique words divided by the number of total words (type-token ratio). Higher type-token ratio indicates more word types in that language. Speaker population data were taken from a variety of sources, where L2 speakers were restricted to adult non-native speakers only. [n = 81]

Moran, McCloy and Wright (2012). Estimates of number of consonants and vowels in each language were used from this dataset. [n = 969]

Lewis and Frank (2014). This work finds that languages tend to map more complex meanings (measured via semantic norms) to longer words. The bias is estimated as the correlation (Pearson's r) between word length and complexity ratings for a set of 499 words translated via Google Translate. We used estimates of the correlation that partialed out the effect of spoken frequency. [n = 79]

Wichmann, Rama, and Holman (2014). This database contains translations for 40-lexical items across many languages. Word length was calculated as the mean number of characters in the ASJPcode transcription system across words in each language. [n = 4421]

Aggregating across datasets, we analyzed 8 environmental variables in total: L2-L1 population ratio, total population size, number of neighbors, area of spoken region, mean

[2]We would like to thank Gary Lupyan and Rick Dale for sharing their data with us.

and standard deviation temperature, and mean and standard deviation precipitation. These variables were selected from a larger set because they were not highly correlated with each other ($r < .8$). We analyzed 6 total linguistic variables: number of vowels, number of consonants, word length, type-token ratio, complexity bias, and morphosyntactic complexity.

5.2.2 Method

Datasets were merged using common ISO-639 codes when available. Five variables were log-transformed to better approximate a normal distribution (population, L2 to L1 ratio, number of neighbors, area, number of consonants, number of vowels).[3]

Main analysis

We tested for a linear relationship between each environmental and language variable. A significant challenge in making inferences about language data is non-independence. This non-independence can come from at least two sources: genetic relatedness and language contact. Following Jaeger et al. (2011), we control for these factors statistically by using linear mixed-effects regression. We control for genetic non-independence by including a random intercept and slope by language family. We control for language contact by including country of origin as a random intercept (models with random slopes failed to converge).[4] We selected country of origin as a proxy for linguistic community because it was available for all languages in our dataset. Both control variables were taken from the WALS dataset. We considered a predictor significant if the test statistic on the fixed effect coefficient exceeded 1.96.

[3] All code and data for the paper are available at `http://github.com/mllewis/langLearnVar`

[4] The model specification was as follows: `language.variable ~ environmental.variable + (environmental.variable | language.family) + (1 | origin.country)`.

Principal component analysis

This first analysis provides a uniform analysis of the many environmental and linguistic variables that have been used to test the Linguistic Niche Hypothesis. However, the large number of variables makes it difficult to distill a coherent picture from these data. Given that many of these variables are partially correlated with each other, we used a technique for reducing the dimensionality of the dataset—principal component analysis. We found the principal components associated with the variance for the environmental variables and the linguistic variables, and then fit the same model as in the primary analysis using the rotated values. Complexity bias was excluded because it was only available for a small subset of languages. All variables were scaled.

5.2.3 Results

In the main analysis, we fit mixed effect models predicting each language variable with each environmental variable using areal controls. The results are presented in Figure 2. For each language variable, there was at least one environmental variable that reliably covaried, though some previously-reported effects were not significant in this analysis. We return to this in the discussion. Data can be explored interactively here: `https://mlewis.shinyapps.io/lhnn/`.

The principal component analysis revealed two primary components of variance for both the environmental and linguistic variables. For the environmental variables, the first two principal components accounted for .69 of the total variance (PC1: .39; PC2: .30). The weights on these variables across the two components can be seen in the upper panel of Fig. 3. The first component loads most heavily on variables related to the climate. It can be thought of as corresponding to hot and rainy regions. The second component loads

most heavily on variables related to the size of the region a language is spoken in, both in terms of number of speakers and physical size. This principal component can be roughly interpreted as the 'smallness' of a linguistic community.

For the linguistic variables, the first two components also accounted for most of the variance, .70 (PC1: .39; PC2: .31; right panel of Fig. 3). The first component loads positively on all variables, except number of vowels. In particular, this component is associated with more consonants, longer words, more word types, and greater morphosyntactic complexity. Broadly, this component is related to the amount of cognitive difficulty associated with learning a language. The second component is associated with having short words, but large phonemic inventories.

Figure 3 shows the relationship between the principal components. Both environmental principal components were reliable predictors of the first linguistic principal component (PC1: $\beta = -0.56$, $t = -3.52$; PC2: $\beta = 0.47$, $t = 2.08$). This suggests that languages that tend to be spoken in cold and small regions are more likely to be more complex. Neither of the environmental principal components were reliable predictors of the second linguistic principal component.

5.2.4 Discussion

These two analyses suggest that more complex languages are spoken in cold, small regions. Importantly, we find this relationship across a range of linguistic features—morphosyntactic complexity, linguistic diversity, word length, and consonant inventory—using the same analytic technique across all measures.

This finding is broadly consistent with previous work that finds relationships between individual metrics of complexity and various demographic variables. Nevertheless, we find

null effects for several reported relationships in the literature. For example, the relationship between population size and morphosyntactic complexity (Lupyan & Dale, 2010) is not reliable in our model with areal controls, though the correlation is significant (r =.08; p <.001) and we replicate their finding in a binned analysis (Fig. 3 of Lupyan & Dale, 2010). There are many possible reasons for these differences (e.g., different measure of complexity, different areal controls), highlighting the need for a common analytical approach across datasets.

Why might languages in small, cold regions have more complex languages? One possible mechanism is that languages spoken in larger places have more L2 learners, and that L2 learners are less skilled than L1 learners at acquiring complex language. As a result, these languages adapt by simplifying. The relationship between climate and linguistic complexity is less clear, but one possibility is that speakers in colder regions are less itinerant, and therefore have less contact with adult speakers of other languages.

5.3 Study 2: Variability in L1 learning

The proposed mechanism in Study 1 makes an important assumption: L2 learners, but not L1 learners, are poor learners of linguistic complexity. Lupyan and Dale (2015) have argued that morphological complexity in fact *facilitates* learning for L1 learners by providing redundancy in the linguistic signal. A straightforward prediction of this hypothesis is that languages that are more easily learnable by L2 learners will be less learnable by L1 learners.

In Study 2, we explore this prediction. As a proxy for language learnability for L1 learners, we use the mean age of acquisition (AoA) of words in a language by L1 learners (children). If there is a tradeoff between learnability for L1 and L2 learners, languages that

are less complex should be harder for children to learn, and thus have later AoAs.

5.3.1 Method

We use subjective measures of AoA from the the Łuniewska et al. (2015) norms. These AoAs were collected from adult participants for the translation equivalents of 299 words in 25 languages. To evaluate the validity of this measure, we compared these ratings to more objective measures of AoA collected from parent-report using the CDI (Wordbank; Frank, Braginsky, Yurovsky, & Marchman, in press). We fit a model predicting the objective ratings with the subjective ratings for the small sample of common languages ($n = 7$). We included language as a random by-intercept and by-slope effect. Subjective ratings were a strong predictor of objective ratings ($\beta = 1.00$, $t = 5.45$), suggesting that the Łuniewska et al. (2015) norms were a reasonable proxy for cross-linguistic AoA.

We averaged across words in the Łuniewska et al. (2015) database to get a mean AoA for each language. We then used the same mixed-effect model as in Study 1 to predict AoAs with each of the linguistic and environmental variables analyzed in Study 1.

5.3.2 Results

Number of consonants positively predicted AoA, suggesting that languages with more consonants tend to have later AoAs ($\beta = 1.04$, $t = 1.97$). In addition, temperature positively predicted AoA ($\beta = .13$, $t = 2.57$) and variability in precipitation negatively predicted AoA ($\beta = -1.6$, $t = -2.83$). No other variables were significant predictors of AoA.

5.3.3 Discussion

Study 2 explores a prediction about a mechanism for the relationship between population size and linguistic complexity: L1 learning is facilitated by complexity in the linguistic signal (via redundancy), but L2 learning is hindered. We find only limited support for this proposal. Of the factors that loaded on "complexity" in the principal component analysis in Study 1, number of consonants was the only reliable predictor of AoA.

Nevertheless, we do find several surprising correlates of AoA—number of consonants, temperature, and precipitation variability—even in this very small sample of languages. The mechanism underlying these relationships is not clear. It could be for example that L1 learners in colder regions have more language input, and therefore earlier AoAs. Or, if we assume that temperature and variability in precipitation are proxies from L2 pressure (as suggested in Study 1), it could be that languages with more L2 pressure have later AoAs, and therefore are harder for L1 learners to acquire. A larger sample of languages will be needed to address these questions.

5.4 Conclusion

Languages vary in many ways across multiple timescales of analysis. Here we suggest that this variability can be accounted for by considering the relationship between these timescales and two types of pressures, those internal and external to the cognitive system. In the present work, we have explored a hypothesis at the language evolution timescale—the Linguistic Niche Hypothesis—which suggests that cross-linguistic variability is the result of different cognitive constraints of learners and environmental pressures.

We contribute to the empirical findings related to this hypothesis by synthesizing previous correlational evidence using common analytical techniques across datasets. Across a range of linguistic and environmental metrics, we find that more complex languages tend to be spoken in smaller, colder regions. We also find evidence that the learnability of a language for L1 learners may be related to aspects of the language (number of consonants) and the environment (temperature and variability in precipitation). Understanding how *both* child and adult learners shape language systems is an important question for future work.

Accounting for variability at the timescale of language evolution is an empirically challenging enterprise. Moving forward, we suggest that a fruitful avenue for progress is holistic descriptions of the empirical landscape, and appeals to processes at multiple timescales of analysis.

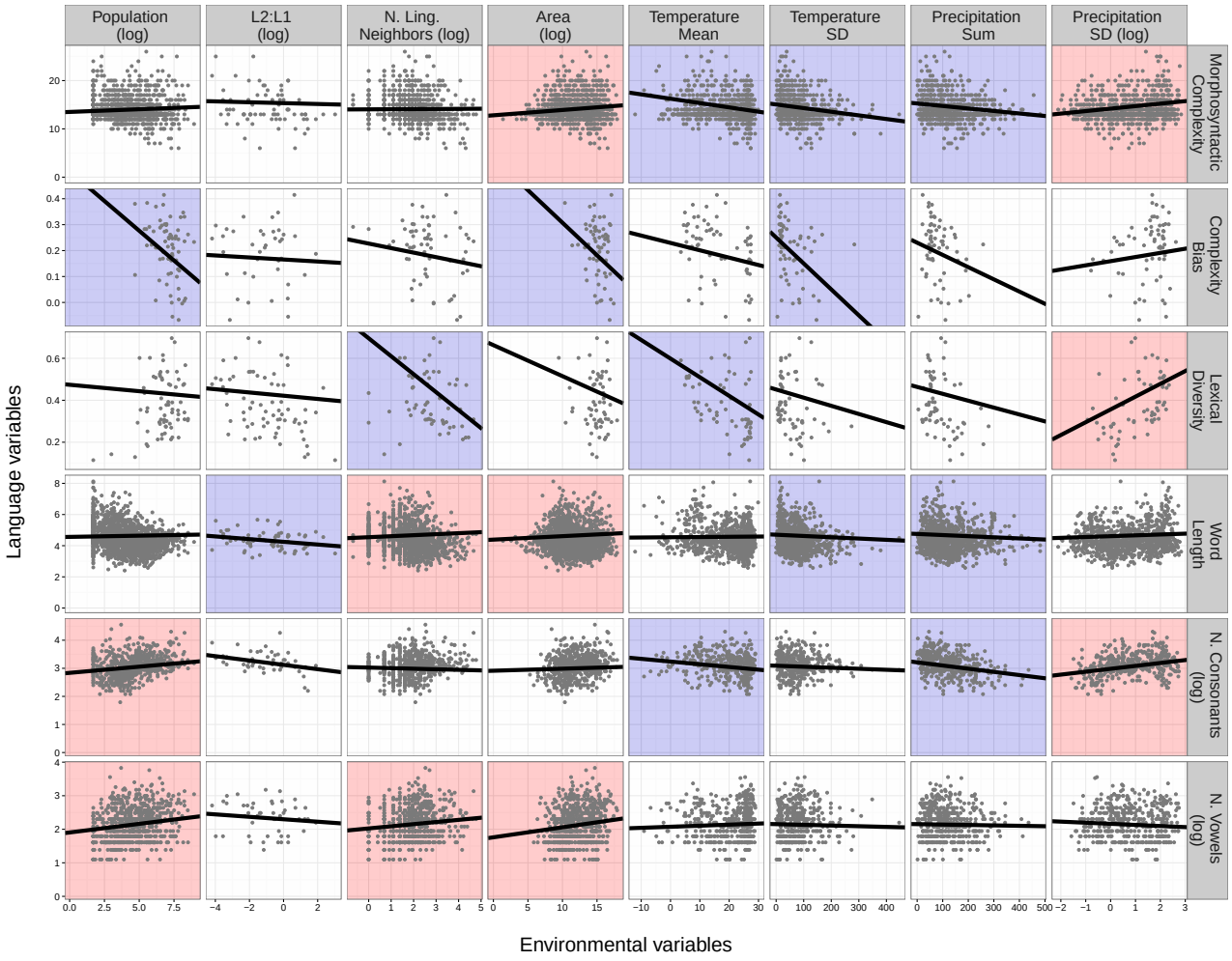

Figure 5.2: Relationship between environmental and linguistic variables, which each point represents a language. Red (positive) and blue (negative) indicate models where the environmental variable is a significant predictor of the linguistic variable. Lines show the fixed effect estimate (slope) and intercept of the mixed effect model. Number of languages varies across plots due to variation in the number of overlapping languages across datasets.

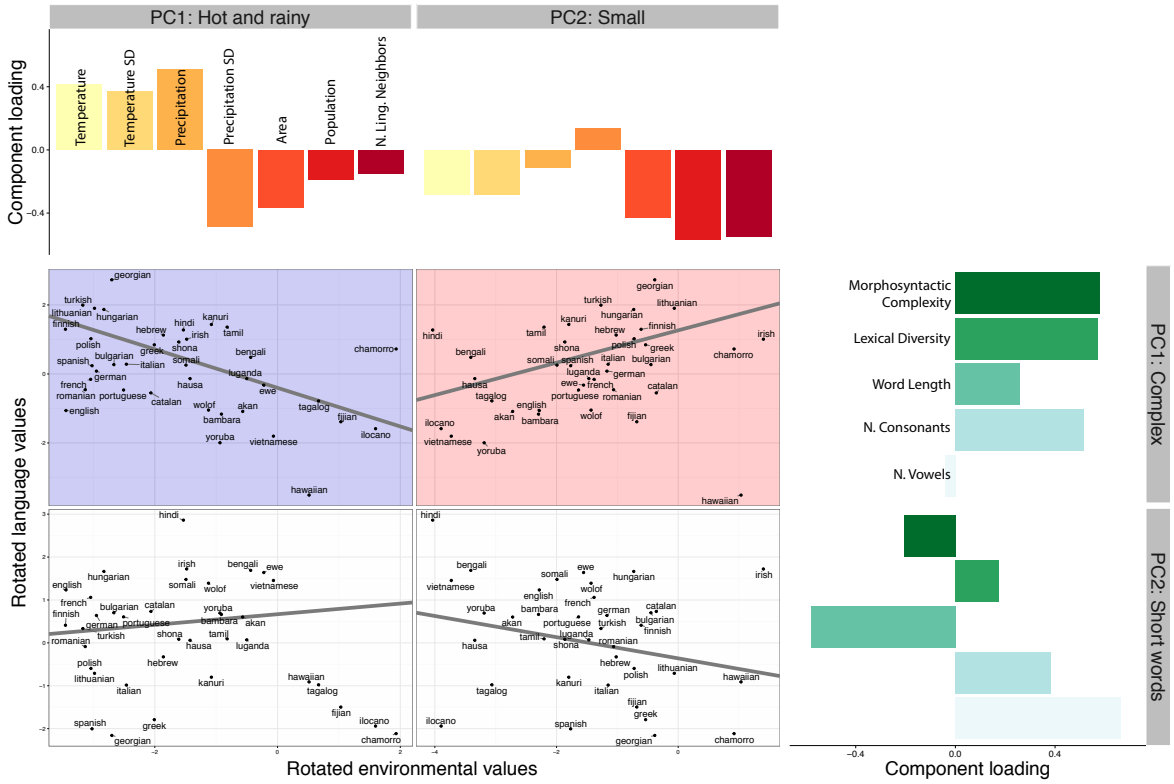

Figure 5.3: Languages spoken in cold, small regions tend to be more complex. The bar plots show the loadings on the first two principal components for the environmental variables (n = 7; orange) and language variables (n = 5; green). The scatter plots show the relationship between the first two principal components for both sets of variables. Each point corresponds to a language, and lines show the linear fit from the mixed effect model. Significance and direction of a linear relationship are indicated by the coloring of the scatterplot (blue: significant and negative; red: significant and positive).

Chapter 6

Conclusion

Important concluding remarks.

Appendix A

A Long Proof

References

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological Science*, *15*(2), 106–111.

Atkinson, Q. D. (2011). Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science*, *332*(6027), 346–349.

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, *47*(1), 31–56.

Baayen, R., Piepenbrock, R., & Gulikers, L. (1995). Celex2, ldc96l14. *Web download, Linguistic Data Consortium, Philadelpha, PA*.

Baddeley, R., & Attewell, D. (2009). The relationship between language and the environment information theory shows why we have only three lightness terms. *Psychological Science*, *20*(9), 1100–1107.

Baldwin, D. (1991). Infants' contribution to the achievement of joint reference. *Child development*, *62*(5), 874–890.

Bentz, C., Verkerk, A., Kiela, D., Hill, F., & Buttery, P. (2015). Adaptive communication: Languages with more non-native speakers have fewer word forms.

Bentz, C., & Winter, B. (2013). Languages with more second language learners tend to

lose nominal case. *Language Dynamics and Change*, *3*(1), 1–27.

Bergen, L., Goodman, N. D., & Levy, R. (2012). Thats what she (could have) said: How alternative utterances affect language use. In *Proceedings of the Thirty-Fourth Annual Conference of the Cognitive Science Society.*

Bergen, L., Levy, R., & Goodman, N. D. (in press). Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*(2), 115.

Blythe, R. A. (2015). Hierarchy of scales in language dynamics.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior Research Methods*, *41*(4), 977–990.

Chater, N., & Christiansen, M. H. (2010). Language acquisition meets language evolution. *Cognitive Science*, *34*(7), 1131–1157.

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, *31*(05), 489-509.

Christiansen, M. H., & Chater, N. (2015). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, 1–52.

Clark, E. (1987). The principle of contrast: A constraint on language acquisition. *Mechanisms of language acquisition. Hillsdale, NJ: Erlbaum*.

Clark, E. (1988). On the logic of contrast. *Journal of Child Language*, *15*(02), 317–335.

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

Clear, J. H. (1993). The British national corpus. In *The digital word* (pp. 163–187).

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210.

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

Dryer, M. S., & Haspelmath, M. (Eds.). (2013). *Wals online.* Leipzig: Max Planck Institute for Evolutionary Anthropology.

Farmer, T. A., Christiansen, M. H., & Monaghan, P. (2006). Phonological typicality influences on-line sentence comprehension. *Proceedings of the National Academy of Sciences*, *103*(32), 12203–12208.

Frank, A., & Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the Twenty-Ninth Annual Conference of the Cognitive Science Society.*

Frank, M., Braginsky, M., Yurovsky, D., & Marchman, V. A. (in press). Wordbank: An open repository for developmental vocabulary data.

Frank, M., Goodman, N., & Tenenbaum, J. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*(5), 578.

Frank, M., & Goodman, N. D. (2014). Inferring word meanings by assuming that speakers are informative. *Cognitive Psychology*, *75*, 80–96.

Frawley, W. (2003). International encyclopedia of linguistics. In (Vol. 2). Oxford University Press.

Genzel, D., & Charniak, E. (2002). Entropy rate constancy in text. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 199–206).

Gordon, R. (2005). *Ethnologue: Languages of the world.* SIL International.

Greenberg, J. (1966). *Universals of language.* Cambridge, MA: MIT Press.

Grice, H. (1975). Logic and conversation. 41–58.

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science*, *31*(3), 441–480.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children.* Paul H Brookes Publishing.

Haspelmath, M. (2006). Against markedness (and what to replace it with). *Journal of Linguistics*, *42*(01), 25–70.

Haspelmath, M. (2008). Frequency vs. iconicity in explaining grammatical asymmetries. *Cognitive Linguistics*, *19*(1), 1–33.

Haspelmath, M., Dryer, M. S., Gil, D., & Comrie, B. (2005). The world atlas of language structures.

Hay, J., & Bauer, L. (2007). Phoneme inventory size and population size. *Language*, *83*(2), 388–400.

Hockett, C. (1960). The origin of speech. *Scientific American*, *203*, 88-96.

Horn, L. (1984). Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. *Meaning, form, and use in context*, *42*.

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.

Hyman, R. (1953). Stimulus information as a determinant of reaction time. *Journal of Experimental Psychology*, *45*(3), 188-196.

Jaeger, T. F., Graff, P., Croft, W., & Pontillo, D. (2011). Mixed effect models for genetic and areal dependencies in linguistic typology. *Linguistic Typology*, *15*(2), 281–320.

Jakobson, R. (1966). Quest for the essence of language. *Morphology, Critical Concepts in Linguistics*, *2004*.

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*(6084), 1049–1054.

Kinzler, K. D., & Spelke, E. S. (2007). Core systems in human cognition. *Progress in Brain Research*, *164*, 257–264.

Kiparsky, P. (1983). Word-formation and the lexicon. In *Proceedings of the 1982 Mid-America Linguistics Conference* (Vol. 3, p. 22).

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Lewis, M., & Frank, M. C. (2016). Learnability pressures influence the encoding of information density in the lexicon learm. In *The Evolution of Language: Proceedings of the 11th International Conference.*

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

Locke, J. (1847). *An essay concerning human understanding.* Troutman & Hayes.

Łuniewska, M., Haman, E., Armon-Lotem, S., Etenkowsk, B., Southwood, F., Pomiechowska, A., Ayiomamitou, B., Boerma, C., Sánchez, C., Dabašinskienė, E., Ehret, E., et al.. (2015). Ratings of age of acquisition of 299 words across 25 languages: Is there a cross-linguistic order of words? *Behavior Research Methods*.

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

Lupyan, G., & Dale, R. (2015). The role of adaptation in understanding linguistic diversity. *The Shaping of Language: The Relationship between the Structures of Languages and their Social, Cultural, Historical, and Natural Environments*.

Mahowald, K., Fedorenko, E., Piantadosi, S., & Gibson, E. (2012). Info/information theory: speakers actively choose shorter words in predictable contexts. *Cognition*, *126*, 313–318.

Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, *126*(2), 313–318.

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

Markman, E., Wasow, J., & Hansen, M. (2003). Use of the mutual exclusivity assumption by young word learners. *Cognitive Psychology*, *47*(3), 241–275.

Maurer, D., Pathman, T., & Mondloch, C. J. (2006). The shape of boubas: Sound–shape correspondences in toddlers and adults. *Developmental Science*, *9*(3), 316–322.

McMurray, B., Horst, J., & Samuelson, L. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review*, *119*(4), 831.

Moran, S., McCloy, D., & Wright, R. (2012). Revisiting population size vs. phoneme inventory size. *Language*, *88*(4), 877–893.

Nettle, D. (2012). Social scale and structural complexity in human languages. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1597), 1829–1836.

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science*, *14*(1), 11–28.

Norvig, P. (2013, February). *English letter frequency counts: Mayzner revisited or etaoin srhldcu.*

Peirce, C. (1931). *The collected papers of Charles S. Peirce.* Cambridge: Harvard University Press.

Piantadosi, S., Tily, H., & Gibson, E. (2011a). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529.

Piantadosi, S., Tily, H., & Gibson, E. (2011b). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280-291.

Piantadosi, S., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280–291.

Reali, 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., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences*, *104*(4), 1436–1441.

Regier, T., Kemp, C., & Kay, P. (2015). Word meanings across languages support efficient communication. In B. MacWhinney & W. O'Grady (Eds.), *The handbook of language emergence* (p. 237-263). Wiley-Blackwell.

Sánchez-Meca, J., Marín-Martínez, F., & Chacón-Moscoso, S. (2003). Effect-size indices for dichotomized outcomes in meta-analysis. *Psychological Methods*, *8*(4), 448.

Saussure, F. (1916, 1960). Course in General Linguistics. London: Peter Owen.

Schmidtke, D. S., Conrad, M., & Jacobs, A. M. (2014). Phonological iconicity. *Frontiers in Psychology*, *5*.

Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, *133*(1), 140–155.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, pp. 379–423.

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

Sutherland, S. L., & Cimpian, A. (2015). An explanatory heuristic gives rise to the belief that words are well suited for their referents. *Cognition*, *143*, 228–240.

Wichmann, S., Rama, T., & Holman, E. W. (2011). Phonological diversity, word length, and population sizes across languages: The ASJP evidence. *Linguistic Typology*, *15*(2), 177–197.

Wierzbicka, A. (1996). *Semantics: Primes and universals.* Oxford University Press, UK.

Wilson, M. (1988). MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, *20*(1), 6–10.

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, Y., & Regier, T. (2014). Numeral systems across languages support efficient communication: From approximate numerosity to recursion. *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*.

Zipf, G. (1936). *The psychobiology of language.* Routledge, London.

Zipf, G. K. (1949). Human behaviour and the principle of least effort. *Cambridge, Mass*.