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 (?, ?)—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 (?, ?, ?). And, despite the growing number of cases suggesting instances of non-arbitrariness in the lexicon (see ?, ?, ?, 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 (?, ?).

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* (?, ?, ?, ?). 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 ? (?), 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., ?, ?). Conceptual primitives can be thought of as the building blocks of meaning, similar to the notion of geons in the study of object recognition (?, ?). 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 (?, ?, ?, ?). 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;* ?, ?, ?). 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 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

[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.

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 ?, ?, 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, ? (?) 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 (?, ?), lightness (?, ?), and numerosity (?, ?).

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

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.

? (?) 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? ? (?) 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.

? (?) 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 (?, ?). 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 (?, ?) and global (?, ?) linguistic context. There is also evidence for this prediction at the level of syntactic (?, ?) and discourse predictability (?, ?).

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. ? (?) 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;" ?, ?). 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 (?, ?). While many notions of iconicity have been discussed in the literature (?, ?, ?), one version of the hypothesis is that linguistic forms often have binary morphemic contrasts and these contrasts map onto a broad difference in meaning (?, ?). 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-"). ? (?) 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. ? (?) 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 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

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) o
9	English complexity norms	2	real words
10	Cross-linguistic corpus analysis	2	real words

Table 2.1: Summary of studies.

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.

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).

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 (?, ?). We created a set of 40 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

[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`.

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/RCind`

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.

2.2 Experiment 2: Mapping Task (Artificial Objects)

2.2.1 Methods

Participants

750 participants completed the experiment.

[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.

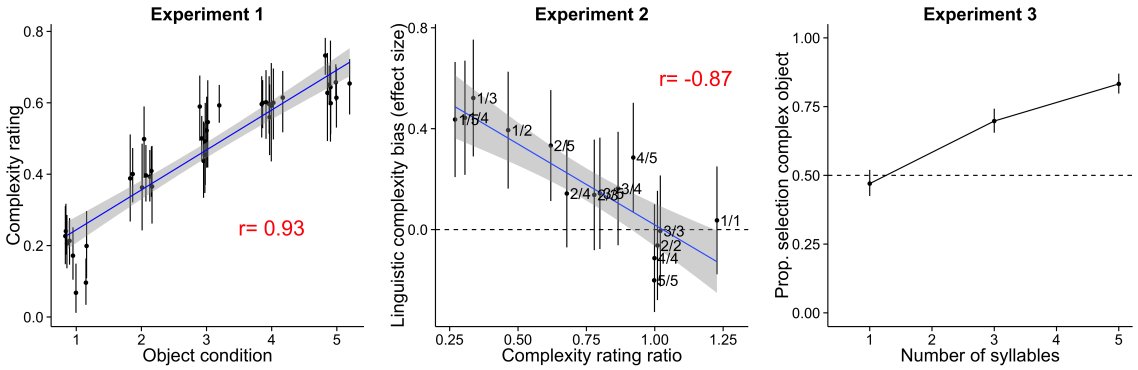

Figure 2.2: (a) The relationship between number of geons and complexity rating is plotted below. Each point corresponds to an object item (8 per condition). The x-coordinates have been jittered to avoid over-plotting. (b) Effect size (bias to select complex alternative in long vs. short word condition) 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.

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 (?, ?). 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.

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).

2.4.1 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.

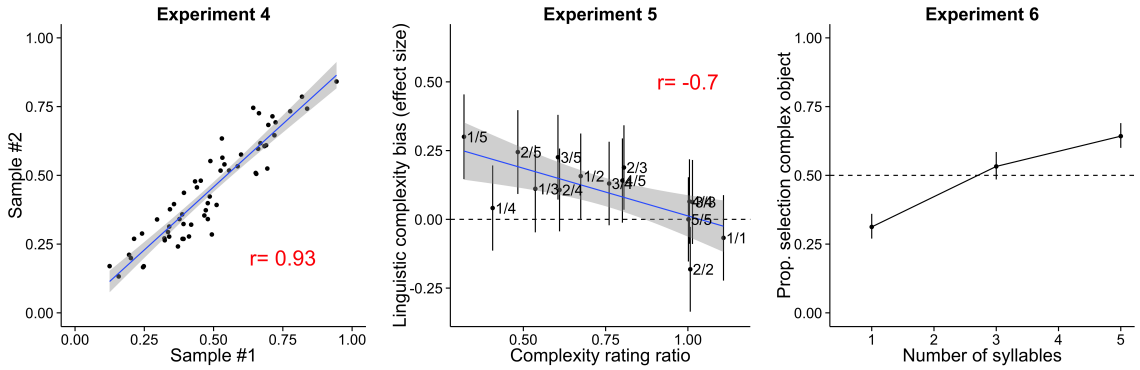

Figure 2.4: (a) The correlation between the two samples of complexity norms. Each point corresponds to an object ($n = 60$). (b) Effect size (bias to select complex alternative in long vs. short word condition) as a function of the complexity rating ratio between the two object alternatives. Each point 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.

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.

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

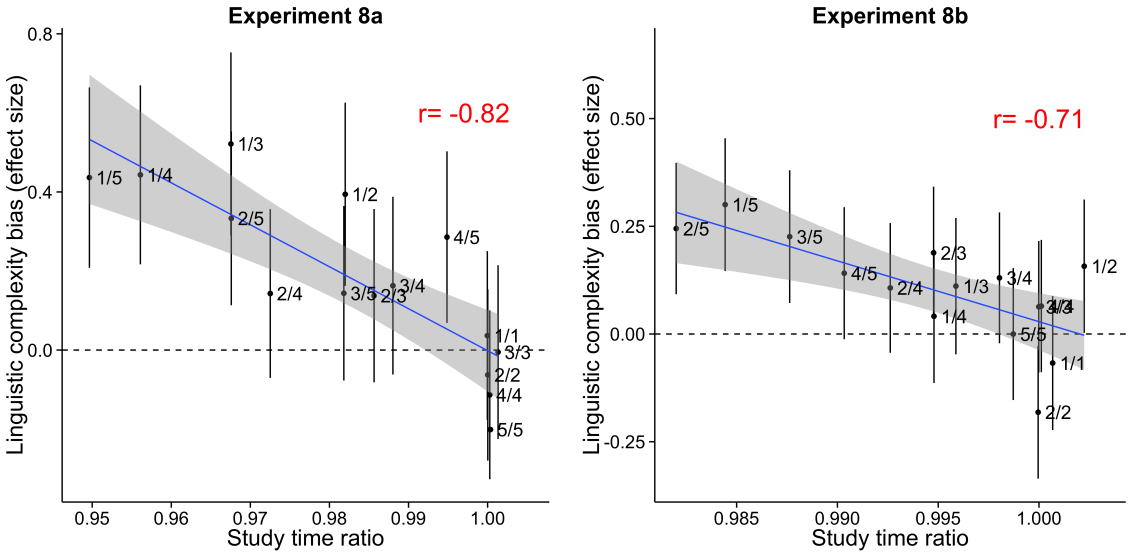

Figure 2.5: Effect sizes in Experiments 2 and 4 replotted in terms of study times collected in Experiment 8. Objects that are studied relatively longer are more likely to be assigned a longer label, relative to a shorter label. Error bars show 95% confidence intervals.

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 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., ?, ?) 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 (?, ?) 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 (?, ?) and measures of morphemes were taken from CELEX2 database (?, ?). 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 (?, ?) or using a language model (?, ?). We estimated word frequency from a corpus of transcripts of American English movies (Subtlex-us database; ?, ?). Importantly, the regularity we

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

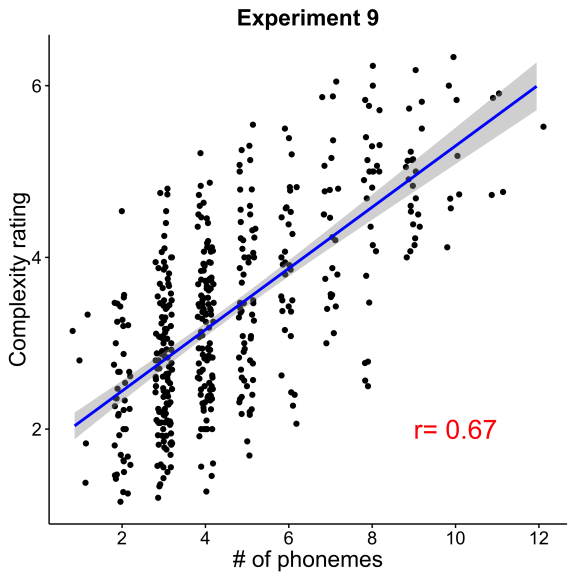

Figure 2.6: Complexity norms collected in Experiment 9 as a function of word length in terms of number of phonemes. Words rated as more complex tend to be longer. Error bars show bootstrapped 95% confidence intervals.

describe—a relationship between conceptual complexity and word length—holds even 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 (?, ?). We included bigram surprisal values for our set of 499 words calculated from the British National Corpus (?, ?).[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

[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.

complexity, frequency, and surprisal, complexity and surprisal were reliable predictors of length ($\beta = 1.11$, $t = 17.22$, $p < .0001$; $\beta = .66$, $t = 2.3$, $p = .02$), 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; ?, ?, $n = 387$; $r_{phonemes} = .53$, $r_{syllables} = .47$, all ps $< .0001$).

Finally, languages also show phonological iconicity effects, such that semantic features (?, ?) and even particular form classes (?, ?) 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 (?, ?). 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; ?, ?). 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.

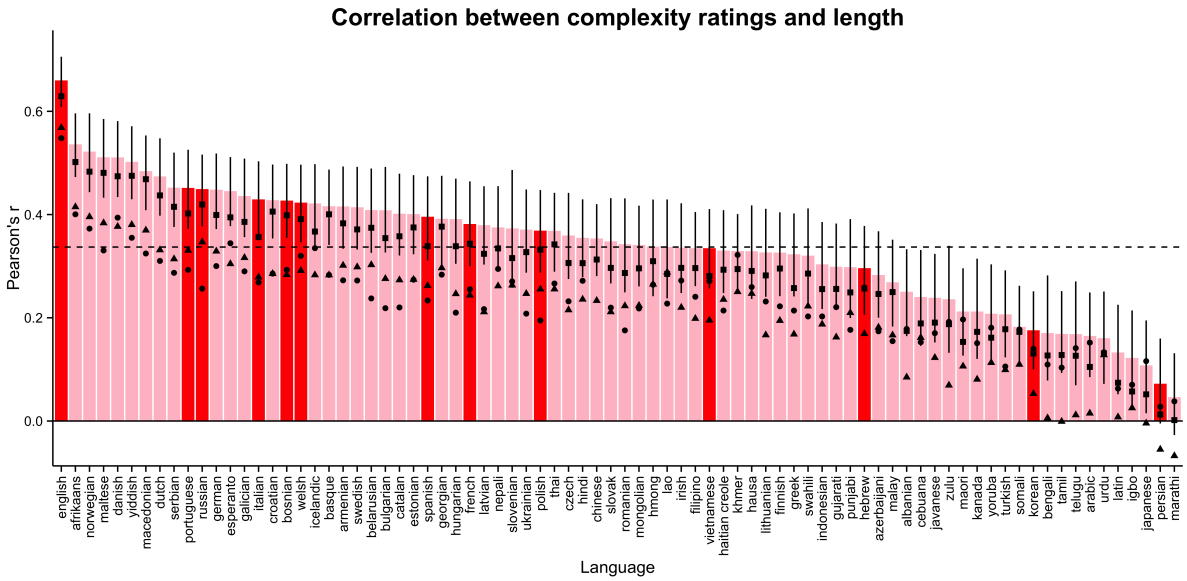

Figure 2.7: Correlation coefficient (Pearson's r) between length in unicode characters and conceptual complexity rating (obtained in Experiment 9). Dark red bars indicate languages for which translations were checked by native speakers; all other bars show translations obtained via Google Translate. 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. In other research, we have explored a number of hypotheses about factors that may contribute to conceptual complexity (see Supplemental Information, Experiments 11 and 12). In particular, we hypothesized that the frequency of objects might contribute to conceptual complexity, such that more frequent objects in the world were less conceptually complex. Across two experiments using similar methods to those reported in the main text, we found no evidence that frequency contributed to complexity. Thus, we leave this difficult topic for future investigations.

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 (?, ?), 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

Introduction

Your introduction here...

Chapter 4

Introduction

Your introduction here...

Chapter 5

Introduction

Your introduction here...

Chapter 6

Introduction

Your introduction here...

Appendix A

A Long Proof

..