The length of words reflects their conceptual complexity

Molly L. Lewis

Department of Psychology, Stanford University

Michael C. Frank

Department of Psychology, Stanford University

We gratefully acknowledge the support of ONR Grant N00014-13-1-0287.

Address all correspondence to Molly L. Lewis, Stanford University, Department of Psychology, Jordan Hall, 450 Serra Mall (Bldg. 420), Stanford, CA, 94305. Phone: 650-721-9270. E-mail: mll@stanford.edu

Abstract

Are the forms of words systematically related to their meaning? The arbitrariness of the sign has long been a foundational part of our understanding of human language. Theories of communication predict a relationship between length and meaning, however: Longer descriptions should be more conceptually complex. Here we show that both the lexicons of human languages and individual speakers encode the relationship between linguistic and conceptual complexity. Experimentally, participants mapped longer words to more complex objects in comprehension and production tasks and across a range of stimuli. Explicit judgments of conceptual complexity were also highly correlated with implicit measures of study time in a memory task, suggesting that complexity is directly related to basic cognitive processes. Observationally, judgements of conceptual complexity for a sample of real words correlate highly with their length across 80 languages, even controlling for frequency, familiarity, imageability, and concreteness. While word lengths are systematically related to usage—both frequency and contextual predictability—our results reveal a systematic relationship with meaning as well. They point to a general regularity in the design of lexicons and reinforce the importance of cognitive constraints on language evolution.

Keywords: communication, lexicon, language evolution

Introduction

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. The 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 (?, ?, for a review), 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).

Importantly, however, the arbitrary property of language is true only from the perspective of the analyst observing a language system from the outside. In contrast, language is highly non-arbitrary from the perspective of a speaker; there are strong constraints on how language is used in particular instances of communication. 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 referential phrase,

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, the utterance "Lee got the car to stop" 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—is a "tupabugorn" more likely to be a complex, unusual object than a "ralex"? In particular, we test the following hypothesis:

> *Complexity Hypothesis*: Languages encode conceptually more complex meanings with longer linguistic forms.

An important construct for our hypothesis is the notion of conceptual complexity. In the present experiments, we study conceptual complexity be manipulating it visually and also measuring it—both directly through explicit norms and indirectly through reaction time. Nonetheless, these metrics serve only as proxies for an underlying cognitive construct. One theoretical framework for understanding this construct is through semantic primitives (e.g. ?, ?). Semantic primitives can be thought of as the building blocks of meaning, similar to the notion of geons in the study of object recognition (?, ?). The space of possible meanings could then be described in terms of sets of semantic primitives. In 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). While our work here does not directly address the character of these underlying semantic primitives, it assumes that such a unit exists and

that meanings can vary in the number of their compositional primitives.

The plan of the paper is as follows. 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. Next, we present nine studies that explore a complexity bias in the lexicon. In Experiments 1-7, we experimentally 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. Finally, we examine a complexity bias in natural language by eliciting complexity norms for English words (Experiment 9) and conducting a corpus analysis of 79 additional languages. We find a robust complexity bias in both novel words and natural language.

Pragmatic equilibria in the lexicon

The present hypothesis is motivated by the possibility that language dynamics take place over different timescales, and these different dynamics may be causally related to each other (?, ?, ?). Minimally, two timescales are relevant to the present hypothesis. At the shorter timescale are the minutes of a single communicative interaction—*the pragmatic timescale*. At the longer timescale is 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. There are other reasons why a regularity like a complexity bias might emerge in the structure of the lexicon, however, and we consider some of these alternative possibilities in the General Discussion.

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 tradeoff with each other: the optimal solution to the speaker's pressure is a single word that can refer to all meanings, while

the optimal solution to the hearer's pressure is a verbose, minimally ambiguous phrase. The utterance that emerges is argued to be an equilibrium between these two tradeoffs.

At the timescale of language evolution, there a number of cases in which these pragmatic equilibria are reflected in the lexicon. One way these equilibria are reflected is in the size of the semantic space denoted by a particular word. From the hearer's perspective, Horn argues there is 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. One example of this is the word "rectangle." This word 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. An example of this is the broadening of brand names to refer to a kind of product. For example, "kleenex" is a name of 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 body of empirical work has tested this prediction by examining the organization of particular semantic domains cross-linguistically (?, ?). Languages show a large degree of similarity in how they partition semantic space for a particular domain, but they also show a large degree of variability. The attested systems can be shown to all approximate an equilibrium point between speaker and hearer pressures.

? (?) 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 one in particular 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 is near the optimal tradeoff between pragmatic pressures. This type of analysis has also been done for the domains of color (?, ?), light (?, ?), and numerosity (?, ?).

A second phenomenon that is predicted by these forces is cases where there are multiple meanings associated with a word from a context-independent perspective, or cases of lexical ambiguity. Lexical ambiguity is present in both open-class words like "bat" (a baseball instrument or a flying mammal) and closed-class quantifiers like "some" ("at least one and possibly all" or "at least one but not all"). 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, Horn’s speaker principle provides a pressure to use the shorter form. This 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. An example of a temporal indexical form is “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”). Language structure reflects this pressure through lexicalized ambiguity in the form of indexicals.

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 hearer is one in which each meaning maps to its own word — exactly a language biased against

synonymy. It turns out that the speaker's pressure also biases against synonymy. The optimal language for the speaker is a language where a single word maps to all meanings. But, a case where multiple words map to a single meaning is also undesirable because the speaker must keep track of two words. So, for both the speaker and the hearer, there is pressure to avoid synonymy. 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. This pattern is reflected in language structure by a one-to-one pattern in the lexicon — that is, a structure in which each word maps to exactly one meaning and each meaning maps to exactly one word.

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 due to the fact that they would have the same meaning as the existing form because they have the same root. 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 commonality 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.

Accounts of language length

Language forms vary along many dimensions, but a salient dimension is length: words and entire utterances can have dramatically different phonetic lengths. ? (?) 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 other pressures at the pragmatic timescale that might influence the length of linguistic expressions. Several theories of communication predict that longer expressions should be associated with less predictable or typical meanings than their shorter counter parts. One such theory is Horn's theory of communication (1984). A speaker often has the choice of using two different utterances to refer to the same meaning (in truth functional terms), and often these utterances differ in length. Horn presents the following example:

(a) Lee stopped the car.

(b) Lee got the car to stop.

Both (a) and (b) have the same denotational meaning (the successful stopping of a car), but they differ in length ((b) has two extra words). Horn argues 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 no only do these two forms differ in meaning, but that they map onto meanings in a systematic way. In particular, he argues that the longer form gets mapped on to the more marked meaning, while the shorter form refers to the unmarked meaning. The notion of 'markedness' is underspecified here, but an intuitive definition is related to complexity: more marked things are more conceptually complex, while less marked things are more conceptual simple. Thus, in the above example, (a) would refer to a simple, average case of car stopping, while (b) 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 of different degrees of markedness 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? This is difficult to derive from models of pragmatic reasoning. ? (?) 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 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 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*. While many notions of markedness have been discussed in the literature (?, ?), one version of the hypothesis is that linguistic forms often have binary contrasts in terms of length and these contrasts map onto a broad difference in meaning—markedness (?, ?). For example, consider the pair "real" - "unreal" :they differ in their valence—positive vs. negative—and the negative form is longer (i.e. has the extra morpheme "un-"). ? (?) claims that this is because there is a linguistic universal that negative meanings are conceptually marked than their positive counterparts. 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, one limitation of this work is that there is no *a priori* criteria for determining what characterizes conceptual markedness; the accounts tend to be just-so stories 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 (e.g. "walk") should be less marked than the past (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, however, in positing conceptual complexity as a broad construct that can be useful applied universally to all meanings.

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 less predictable meanings, but no work directly tests this prediction. This was the goal of the present studies.

Complexity bias in comprehension: Artificial objects

As a first step in exploring a complexity bias, we manipulated the complexity of objects and asked participants to infer their names. 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 robust complexity bias. Experiment 3 replicated Experiment 2 with randomly concatenated syllables.

Experiment 1: Object complexity norms

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 1a and 1b with geons).

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

Figure 1. Stimuli 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).

containing 1-5 geon primitives (Figure 1)[1]

Procedure. We presented participants 12 objects from the 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. The task can be viewed directly here: `http://langcog.stanford.edu/expts/MLL/refComplex/Experiment34/ref_complex_34.html`

Results and Discussion

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

[1] All stimuli, experiments, raw data and analysis code can be found at `https://github.com/mllewis/RC`. Analyses can be found at: `http://rpubs.com/mll/50311`.

for manipulations of conceptual complexity.

Experiment 2: Mapping task

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 (e.g., "bugorn" and "tupabugorn") long. 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."; `http://langcog.stanford.edu/expts/MLL/refComplex/Experiment38/ref_complex_38.html`).

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.

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 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.

Experiment 3: Control mapping task

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 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.

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 2, 4, or 6 consonant-vowel syllables (e.g., "nur," "nobimup," "gugotobanid"). The last syllable of all words ended in a consonant to better approximate the phonology of English.

Procedure. Participants completed six forced-choice trials identical to Experiment 1b (`http://langcog.stanford.edu/expts/MLL/refComplex/Experiment40/ref_complex_40.html`). 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 (2, 4, or 6 syllables).

Results and Discussion

Replicating the "1/5" condition in Experiment 1b, 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 ($\beta = -.44$, $p < .0001$). This suggests that the complexity bias observed in Experiment 2 is unlikely to be due to the particular set of words we selected.

Complexity bias in comprehension: Novel real 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 instead of artificial geon objects. We find that the complexity bias observed with geon objects extends to naturalistic objects.

Experiment 4: Object complexity norms

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 did not have canonical labels associated with them (Figure 2).

Procedure. The procedure was identical to Experiment 1 (`http://langcog.stanford.edu/expts/MLL/refComplex/Experiment9/ref_complex_9.html`).

Results and Discussion

Complexity judgements were highly reliable across two independent samples ($r = .93, p < .0001$; $M_1 = .49$, $SD_1 = .18, M_2 = .44$, $SD_2 = .18$). Figure 3a shows the relationship

Figure 2. Stimuli in Experiments 4-6: naturalistic objects without canonical labels. Each row corresponds to a quintile determined by the explicit complexity judgements obtained in Experiment 4 (top: least complex; bottom: most complex).

between the complexity judgment for each item across the two samples of participants.

Experiment 5: Mapping task

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 2a. 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 (`http://langcog.stanford.edu/expts/MLL/refComplex/Experiment35/ref_complex_35.html`).

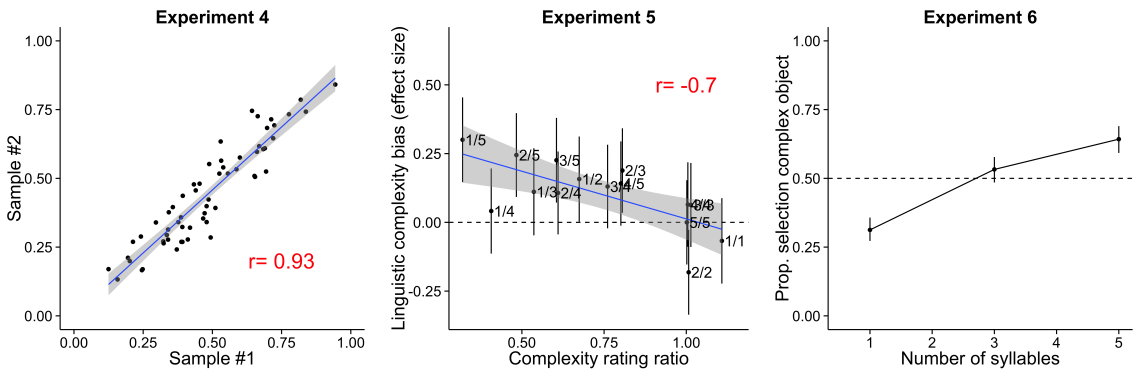

Figure 3. (a) The correlation between the two samples of complexity norms. Each point corresponds to an object ($n = 60$). (b) Effect size (bias to select complex alternative in long vs. short word condition) as a function of the complexity rating ratio between the two object alternatives. Each point 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.

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. 3b). 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 on natural language more generally.

The effect size in Experiment 5 is smaller than in Experiment 2, however. This 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 robust complexity bias

with naturalistic objects.

Experiment 6: Control mapping task

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 (`http://langcog.stanford.edu/expts/MLL/refComplex/Experiment41/ref_complex_41.html`).

Results and Discussion

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 ($\beta = -.34, p < .0001$; Fig. 3c). 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 (Experiment 5; STATS). This 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."). This may 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 may be over all less conceptually 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.

Complexity bias in production

The previous set of experiments provide evidence for a complexity bias in a comprehension task with novel words. One limitation of the design, however, is the possible presence of 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.

Experiment 7: Production task

Methods

Participants. Fifty-nine participants completed the experiment.

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

Procedure. In each trial, we presented with 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" (`http://langcog.stanford.edu/expts/MLL/refComplex/Experiment27/ref_complex_27.html`). Each participant completed 10 trials—five objects from the bottom and top complexity norm quantiles each. Order of objects was randomized.

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 ($M = 1.94$, $SD = 0.18$) compared to the bottom quartile ($M = 1.85$, $SD = 0.17$; $t(57) = 3.92$, $p < .001$). We also analyzed length as a function of the complexity norms for each object. Length of production was correlated with the complexity norms: Longer labels were coined for objects that were rated as more complex ($r = .17, p < .0001$). This experiment provides a strong test of the complexity bias: Even with minimal task demands, participants prefer to use longer words to refer to more complex objects.

Complexity as a cognitive construct

[these two sentences are from the cogsci paper] In Study 3, we try to more directly examine the cognitive correlates of conceptual complexity through reaction time.

Experiments 1-7 suggest that participants have a productive complexity bias when complexity is operationalized in terms of explicit norms. Next we sought to explore the cognitive construct of 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. In visual cognition, stimuli that contain more information require more processing time in search (?, ?, ?). To test this prediction, 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.

Methods

Participants. 750 participants completed the task. 250 participants were tested with artificial geon objects (Experiment 3a) and 500 were tested with naturalistic objects (Experiment 3b).

Stimuli. The study objects were the set of 40 artificial geon objects (Experiment 8a) and 60 naturalistic 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 geon objects and 30 naturalistic 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 geon object and 60 naturalistic objects), and asked "Have you seen this object before?." Participants responded by clicking a "yes" or "no" button (8a: `http://langcog.stanford.edu/expts/MLL/refComplex/Experiment37/ref_complex_37.html`, and 8b: `http://langcog.stanford.edu/expts/MLL/refComplex/Experiment30/ref_complex_30.html`).

Results and Discussion

Experiment 8a: Geon 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: $M = 7.40$, $SD = .66$).

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.

Study times did not predict memory performance. The study times for hits (correct "yes" responses; $M = 7.33, SD = .52$) did not differ from misses (correct "no" responses; $M = 7.34$, $SD = .59$; $t(223) = .61$, $p = .54$).

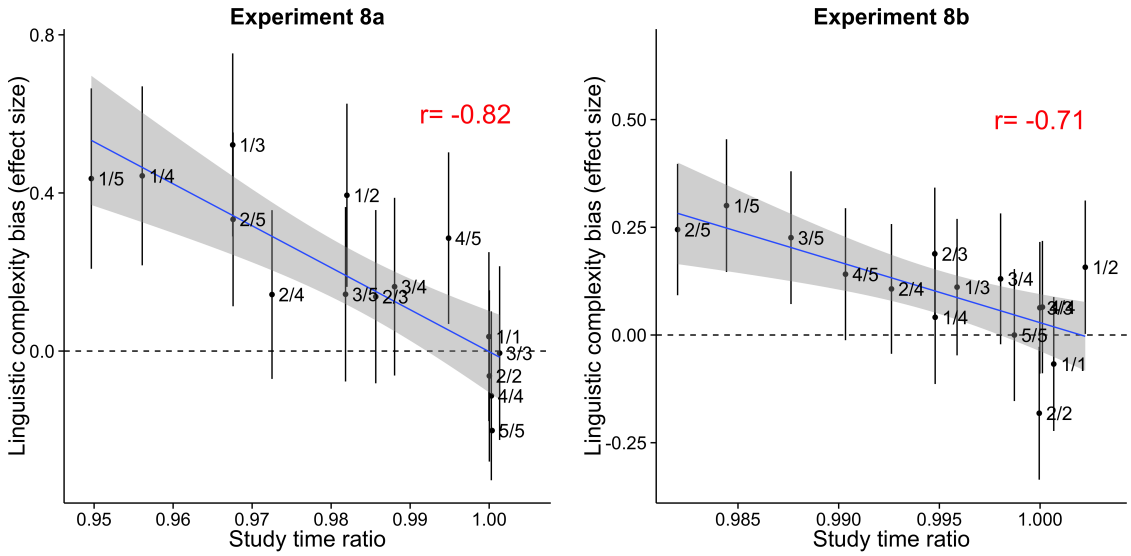

Figure 4. Effect sizes in Experiments 2 and 4 replotted in terms of study times collected in Experiment 8. Objects that are studied relatively longer are more likely to be assigned a longer label, relative to a shorter label. Error bars show 95% confidence intervals.

The critical question was whether or not 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. 4a): Relatively longer study times predicted longer labels.

Experiment 8b: Naturalistic objects. We excluded 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. This excluded 6 participants (1%). 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. This led to the exclusion of 4% of responses (final sample:

$M = 7.36$, $SD = .72$).

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 ($r = .54$, $p < .0001$).

Unlike for the geons, study times predicted memory performance. Study times for hits (correct "yes" responses; $M = 7.24$, $SD = .60$) were greater than for misses (correct "no responses; $M = 7.11$, $SD = .66$; $t(393) = 9.74$, $p < .0001$).

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. 4b).

Together, these findings suggest that label judgments are supported by basic cognitive processes related to the complexity or information content of a stimulus.

Together, these experiments 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.

Complexity bias in natural language

Experiments 1-8 revealed a productive complexity bias in the case of novel words. Next we ask whether this bias extends to natural language. In Experiment 9, we collected explicit complexity judgements 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. We then ask whether these complexity ratings correlate with word length in a sample of 79 languages. We find a complexity bias in all 79 languages, suggesting that this bias is a pervasive property of natural language.

To measure conceptual complexity in natural language, we adopt a rating scale approach similar to that used in previous work to quantify other aspects meaning like how perceptible a

referent is (concreteness) and how much experience speakers tend to have with a referent (familarity; ?, ?) . In this work, participants are presented with a 5- or 7- point Likert scale anchored at both ends of the dimension of interest, and asked to make an explicit judgement about a word's meaning. This approach is not ideal because it requires that all participants conceptualize the dimension of interest in a similar way. Nonetheless, previous work has shown thesE measures to be easy to handle analytically and reliable, and so we adopt them here to quantify conceptual complexity.

Experiment 9: English complexity norms

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 allowed us to compare our complexity norms to previously measured psycholinguistic variables.

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 words meaning is simple if it is easy to understand and has few parts. An example of a simple meaning is "brick." A words 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 (`http://langcog.stanford.edu/expts/MLL/refComplex/Experiment26/ref_complex_26.html`)

Results and Discussion

We considered 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 Experiment 4b below (phonemes: $r = .92$; morphemes: $r = .69$; syllables: $r = .87$).

Given these measures of word length, we considered how length related to judgements of meaning complexity. We excluded participants who missed a simple math problem in the middle of the task that served as an attentional check. This excluded 6 participants (2%). Critically, we found that complexity ratings ($M = 3.36$, $SD = 1.93$) 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. 5). 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.

Word length is strongly related to linguistic predictability, operationalized via simple frequency (?, ?) or using a language model (?, ?). But the regularity we 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 ($r = .57$, $p < .0001$). Word frequency was estimated from a corpus of transcripts of American English movies (Subtlex-us database; ?, ?).

Complexity and length are intuitively related to a number of other psycholinguistic

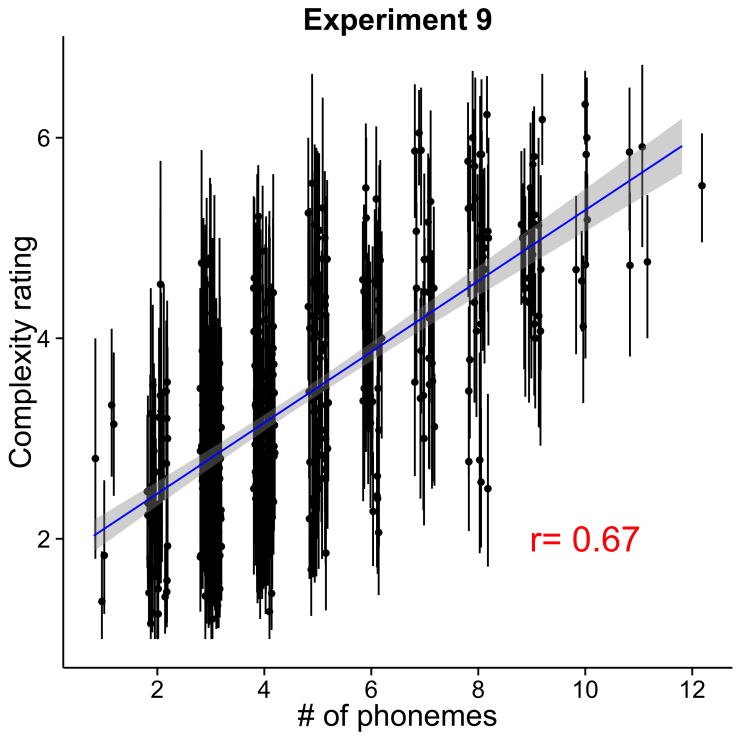

Figure 5. Complexity norms collected in Experiment 9 as a function of word length in terms of number of phonemes. Words rated as more complex tend to be longer. Error bars show bootstrapped 95% confidence intervals.

variables. All of these variables were reliably correlated with complexity (concreteness: $r = -.27$; familiarity: $r = -.43$; imageability: $r = -.21$). Nonetheless, the relationship between word length and complexity remained reliable controlling for all of 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 below. This pattern held for the other two metrics of word length (morphemes and syllables).

? (?) noted that some forms are more complex, or *marked*, where markedness is denoted by morphological structure. For example, on this account, plurals are considered more complex than

singulars because they are more marked morphologically (by the -s morpheme). 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 $< .0001$).

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 below the median in concreteness, those words whose part structure is presumably much less obvious ($r_{phonemes} = .73$, $r_{syllables} = .72$, $r_{morphemes} = .47$, all ps $< .0001$).

*direction of causality (long -¿ rated as more complex)

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 across all the languages we examined.

Methods

Results and Discussion

General Discussion

alternative accounts of complexity bias * iconicity * naming hypothesis

need more direct evidence for a causal link

what is complexity??

References