The length of words reflects their conceptual complexity

Molly L. Lewis and Michael C. Frank

Psychology Department, Stanford University,
450 Serra Mall, Stanford, CA 94305, USA

*To whom correspondence should be addressed; E-mail: mll@stanford.edu.

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.[1,2] Theories of communication predict a relationship between length and meaning, however: Longer descriptions should be more conceptually complex.[3,4] 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 frequency and contextual predictability,[5,6] our results reveal a systematic relationship between word length and meaning. They point to a general regularity in the design of lexicons and suggest the importance of communicative and

cognitive constraints on language evolution.[7,8]

In a classic example of pragmatic reasoning from Horn,[3] 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. Do we reason the same way about the meanings of words? Is a "tupabugorn" more likely to be a complex, unusual object than a "ralex"?

We tested this hypothesis by asking whether speakers would be biased to interpret a long novel word as being more likely to refer to a more complex novel referent. We presented participants on Amazon Mechanical Turk with a novel word of either 2 or 4 syllables and two possible objects as referents (Study 1: $n = 750$; Fig. 1a). Possible referents were novel artificial objects whose complexity we manipulated by varying the number of parts the object contained (1 - 5 "geons";[9] Fig. 1b; these judgements were highly correlated with explicit complexity judgments, Study 2: $n = 60$, $r = .93$, $p < .0001$). Participants were asked to select which object the word named for every unique combination of object complexities (1 vs. 2 geons, 1 vs. 3 geons, 1 vs. 4 geons, etc.).

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 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 ($r = -.87$, $p < .0001$; Fig. 1c). In a control experiment, we also found this bias with words that were composed of randomly concatenated syllables: participants were more likely to select a five geon object

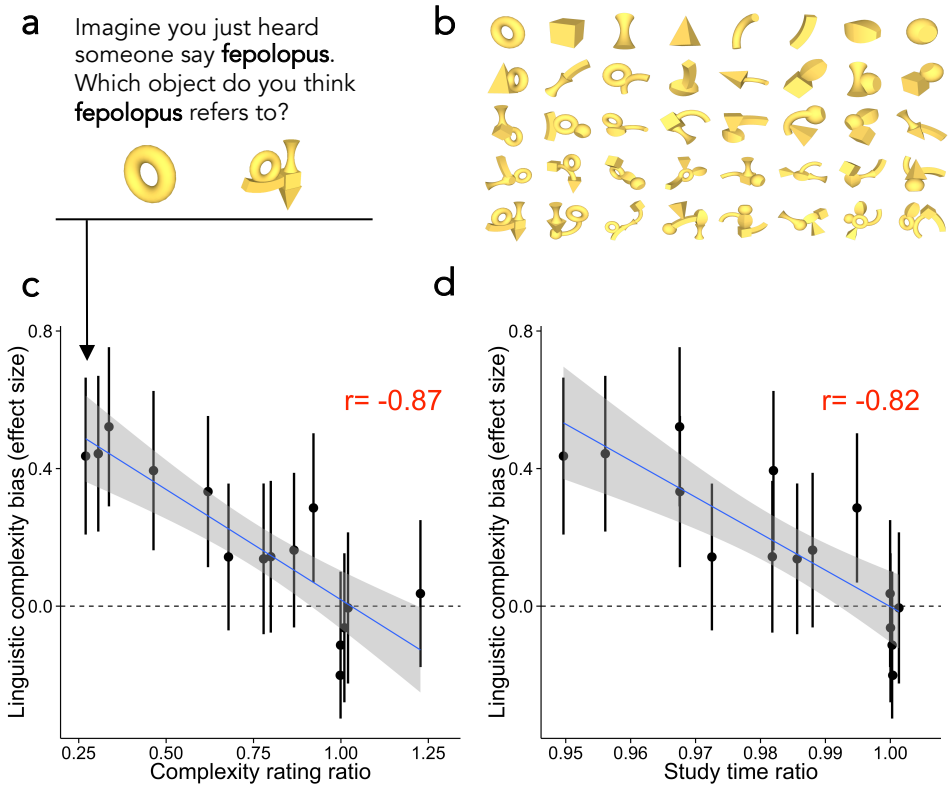

Figure 1: (a) Schema of a 1 vs. 5 geon trial and the corresponding data point. One referential alternative contains one geon and the other contains five geons. (b) Artificial "geon" stimuli. Each row shows a different level of complexity, determined by the number of geon parts in the objects. (c, d) Experimental results from a task in which participants were asked to map a novel word of varying length to one of two possible referents ($n = 750$). Effect size between the long and short language conditions is plotted against the complexity ratio of the two referent alternatives. Fig. 1c shows the referents plotted in terms of explicit complexity judgements, and Fig. 1d shows the referents plotted in terms of study time. Effect sizes were calculated using the log odds ratio (see Supplementary Information for further details). Error bars show 95% confidence intervals.

compared to a single geon object as the number of syllables in the word increased (Study 3: $n = 200$; $\beta = -.44$, $p < .0001$).

Next we asked whether this bias extended to more naturalistic objects. We gathered a sample of real objects without canonical labels (Fig. 2a) and asked participants to rate their complexity.

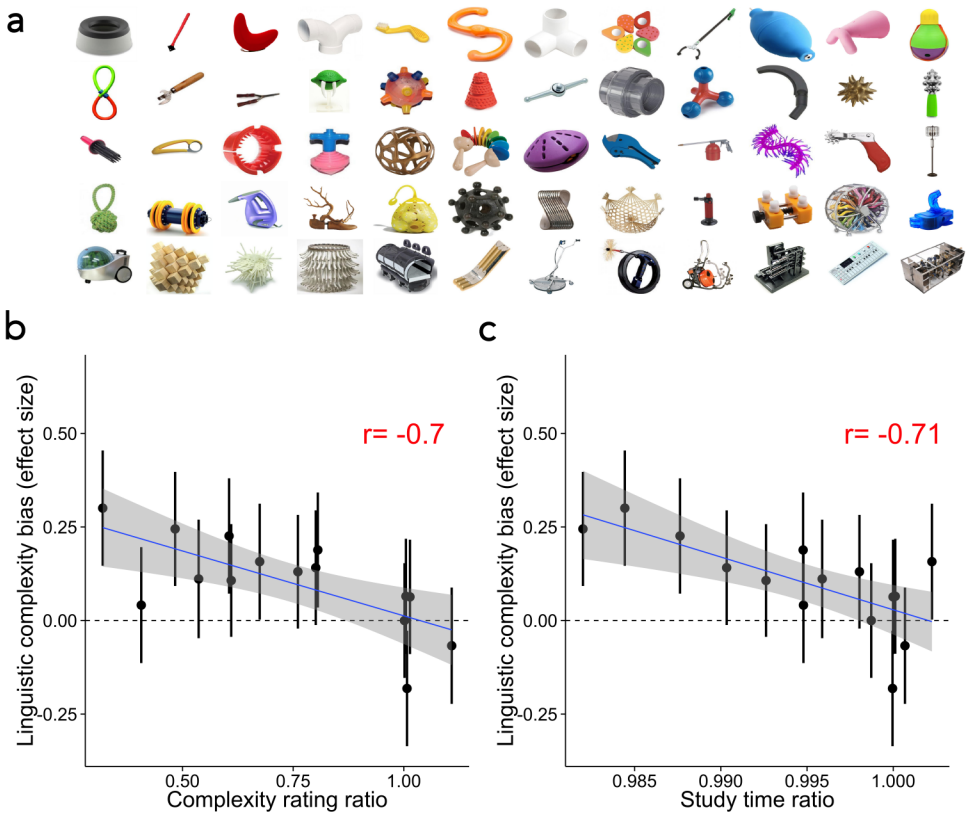

Figure 2: (a) Naturalistic, novel stimuli. Each row corresponds to a quintile determined by the explicit complexity judgements (top: least complex; bottom: most complex). (b, c) Experimental results from a task in which participants were asked to map a novel word of varying length to one of two possible referents ($n = 1500$). Effect size between the long and short language conditions is plotted against the complexity ratio of the two referent alternatives. Fig. 2b shows the referents plotted in terms of explicit complexity judgements, and Fig. 2c shows the referents plotted in terms of study time. Effect sizes were calculated using the log odds ratio (see Supplementary Information for further details). Error bars show 95% confidence intervals.

These judgements were highly reliable across two independent samples (Study 4: $n = 60$ in each, $r = .93$, $p < .0001$). We then divided the objects into quintiles based on these ratings, and used them as stimuli in a mapping task identical to the one used with the artificial objects. As with the artificial objects, effect size was negatively correlated with the complexity rating ratio between the referent alternatives (Study 5: $n = 1500$; $r = .70$, $p < .005$; Fig. 2b). We also

replicated this result with randomly concatenated syllables, such that 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 (Study 6: $n = 200$, $\beta = -.34$, $p < .0001$). Finally, we found the same bias in language production: Participants produced novel coinages that were longer for the top quartile of objects compared to the bottom quartile (Study 7: $n = 59$; $t(57) = 3.92$, $p < .001$).

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.[10,11] 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.

Mean study time was highly correlated with explicit complexity norms for both artificial objects (Study 8: $n = 250$; $r = .89$, $p < .0001$) and novel real objects (Study 9: $n = 500$; $r = .54$, $p < .0001$). In addition, the ratio of study times for the two object alternatives was correlated with the bias to choose a longer label for both the artificial objects ($r = .82$, $p < .001$; Fig. 1c) and the novel real objects ($r = .71$, $p < .005$; Fig. 2c): Relatively longer study times predicted longer labels. 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. Is this bias only relevant to judgments of unfamiliar words, or does it apply to familiar labels as well?

We collected ratings of meaning complexity for 499 English words in a rating procedure

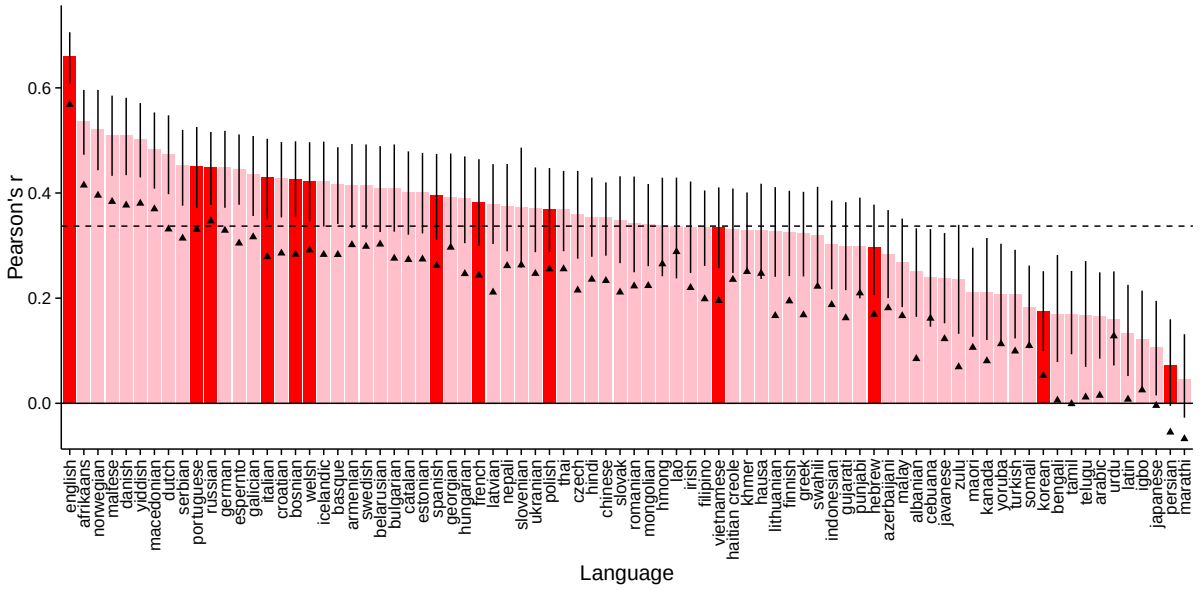

Figure 3: Correlations between conceptual complexity norms and word lengths, across languages. Dark red bars indicate languages for which translations were checked by native speakers; all other bars show translations obtained via Google Translate. Error bars show 95% confidence intervals obtained via non-parametric bootstrap. Triangles indicate correlation value partialling out English log frequency. The dashed line indicates the grand mean correlation across languages.

similar to the objects in Studies 2 and 4 (Study 10: $n = 246$). Complexity judgements were positively correlated with word length, measured in phonemes, syllables, and morphemes ($r_{phonemes} = .67$, $r_{syllables} = .63$, $r_{morphemes} = .43$, all ps $< .0001$), even when closed-class words were excluded ($n = 438$; $r_{phonemes} = .65$, $r_{syllables} = .63$, $r_{morphemes} = .42$, all ps $< .0001$). Importantly, these relationships also remained reliable after controlling for the word's concreteness, imageability, and familiarity.

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, using Google Translate to translate our word set (Study 11). Native speakers checked the accuracy of these translations for 12 of the 79 languages, finding an accuracy of .92 within this sample. For each

language, we calculated the correlation between word length in terms of number of characters (to allow comparison between languages for which no phonetic dictionary was available) and mean complexity rating. All 79 languages showed a positive correlation between length and complexity ratings (Fig. 3). The grand mean correlation across languages was .34.

Word length is strongly related to linguistic predictability, operationalized via simple frequency[5] or using a language model.[6] 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 ($r = .57$, $p < .0001$). Across languages, partialling out log frequency (estimated in English), the grand mean correlation was .22. In addition, when we manipulated the observed frequencies of novel objects experimentally, we found no effects on judgments of word length (Studies 12–13).

Languages also show phonological iconicity effects, such that semantic features[12] and even particular form classes[13] 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 $< .0001$).

Greenberg[14] 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. First,

if participants' conceptual complexity judgements were based on English morphological complexity, we should not expect to see correlations between those judgments and word length in other, unrelated languages like Basque or Mandarin. Second, in English, the correlations we observed hold for words with no obvious derivational morphology (CELEX2 monomorphemes,[15] $n = 387$; $r_{phonemes} = .53$, $r_{syllables} = .47$, all ps $< .0001$).

A purely arbitrary signaling system is inherently symmetric: There is no reason to associate one message with one meaning. The regularity that Horn[3] described—that longer messages have more complex or unusual meanings—provides a method for breaking this symmetry. The same symmetry problem is present in the lexicons of natural languages; our work here suggests that the same regularity applies as well. Our data do not speak to the processes underlying participants' judgments that longer words have more complex meanings, however. In particular, these judgments need not reflect in-the-moment pragmatic inference; they could also be an iconic mapping between effort and meaning or a lower-level statistical regularity extracted through extensive experience with a language. Regardless of its cognitive instantiation, the result is a lexicon that reflects Horn's principle.

Methods In Studies 1 and 5, we manipulated word length (2 vs. 4 syllables) and the relative complexity of the referent alternatives within participants. There were 15 complexity conditions, corresponding to every possible combination of object quintiles. In Study 1, the quintiles were determined by the number of geons in the object. In Study 5, the quintiles were determined by the norms obtained in Study 4. 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.

In Studies 2 and 4, 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." The first two objects were images of a ball and a motherboard to anchor participants on the scale.

In the Studies 3 and 6, participants completed six forced-choice trials in which they saw two possible referents, from the top and bottom quintiles. The novel words were created by randomly concatenating 2, 4, or 6 consonant-vowel syllables. The last syllable of all words ended in a consonant. Each participant completed 2 trials for each word length.

In Study 7, participants were presented 10 objects from the set of objects normed in Study 4 and asked to generate a novel single-word label for the object. Five of the objects were from the bottom quantile of complexity norms, and 5 of the objects were from the top quantile of complexity norms. Order of objects was randomized.

In Studies 8 and 9, 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.

In Study 10, we selected 499 relatively high-frequency English words. For each word, we 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 32 words.

In Study 12 ($n = 477$), we presented participants with 10 objects on a single screen. The objects were composed of a single geon. There were two types of objects. One object type

appeared nine times and the second object type appeared once. After this training period, participants completed a forced choice mapping task, as in Studies 1 and 5. We presented a word that was either 2 or 4 syllables long and asked participants to make a judgment about whether the word referred to the low or high frequency object. Each participant completed a single mapping trial, and word length was manipulated between participants. There was no difference between the long and short word conditions ($\chi^2(1) = 0.02$, $p = .89$).

In Study 13 ($n = 97$), we manipulated object frequency by sequentially presenting objects. Participants saw 60 objects from theset of normed real objects one at a time. One object was presented 10 times and a second object was presented 40 times. Ten additional objects were included as fillers. After this training phase, participants completed a single mapping trial as in Study 12. Word length was manipulated between participants. There was no difference between the long and short word conditions ($\chi^2(1) = 0.01$, $p = .92$).

The Stanford University Review Board approved the study protocol for all experiments, and informed consent was obtained from participants prior to their participation. The sample sizes and exclusion criteria were pre-specified on the basis of pilot studies. Exclusion criteria are described in the Supplementary Information. The data meet the assumptions of the statistical tests applied (see the Supplementary Information for standard deviations), and all statistical tests were two-tailed.

References

1. Saussure, F. In *Course in General Linguistics* (Peter Owen, London, 1916, 1960).

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

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

4. Levy, R. P. & Jaeger, T. F. Speakers optimize information density through syntactic reduction. In Schlökopf, B., Platt, J. & Hoffman, T. (eds.) *Advances in Neural Information Processing Systems*, vol. 19, 849–856 (MIT Press, Cambridge, MA, 2006).

5. Zipf, G. *The Psychobiology of Language* (Routledge, London, 1936).

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

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

8. Lieberman, E., Michel, J.-B., Jackson, J., Tang, T. & Nowak, M. A. Quantifying the evolutionary dynamics of language. *Nature* **449**, 713–716 (2007).

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

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

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

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

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

14. Greenberg, J. *Universals of Language.* (MIT Press, Cambridge, MA, 1966).

15. Baayen, R., Piepenbrock, R. & Gulikers, L. CELEX2, LDC96L14. Web download. In *Linguistic Data Consortium* (Philadelpha, PA, 1995).

Supplementary Information

Supplementary Information can be found at `http://rpubs.com/mll/33927`.

Acknowledgements

Author Contributions

MLL and MCF designed research. MLL conducted research. MLL and MCF wrote the paper.

Author Information