

# LEARNABILITY PRESSURES INFLUENCE THE ENCODING OF INFORMATION DENSITY IN THE LEXICON

ANONYMOUS AUTHOR 1, ANONYMOUS AUTHOR 2

*University Department, University Name*  
*City, Country*  
*email1@university, email2@university*

A universal feature of language is that words vary in their length within a language, in terms of morphemes, syllables, and phonemes. There have been several accounts of this variability in the literature that appeal to the form of language itself, such as the frequency of a word, or its predictability in context. Information theory, however, suggests another factor might influence length: the predictability, or complexity, of a word's meaning (Frank & Jaeger, 2008). This theory predicts that if speakers try to maintain a constant rate of information across the speech stream, then more complex meanings should be longer.

Previous work has shown evidence for this *complexity bias*. Experimentally, participants tend to assign longer words to more complex meanings, relative to shorter words (Lewis, Sugarman, & Frank, 2014). And natural languages also show this bias (Lewis & Frank, under review). To estimate the bias in each language, ratings of conceptual complexity were collected for 499 English words and then translated into 79 additional languages using the Google Translate. For each language, there was a correlation between word length and conceptual complexity (grand mean  $r = .34$ ), and this bias held controlling for word frequency (grand mean  $r = .22$ ) and other semantic variables, like concreteness. Despite the presence of this bias across all languages we examined, however, there was also a large degree of variability ( $SD = .12$ ; Fig. 1).

In our work here, we explore one possible account of this variability: that the degree to which a language encodes conceptual complexity in the lexicon is related to the learnability pressure on the language. Learnability pressure has been argued as one factor influencing the morphological complexity of a language (Lupyan & Dale, 2010). Under this hypothesis, languages are thought to adapt to their particular social context, depending on the cohesiveness of the population acquiring the language. Languages that are acquired by a diverse population of speakers—many adult learners, for example—might be morphologically simpler, than those acquired only by children. Consistent with this prediction, languages that are spoken by more people tend to be less morphologically complex.

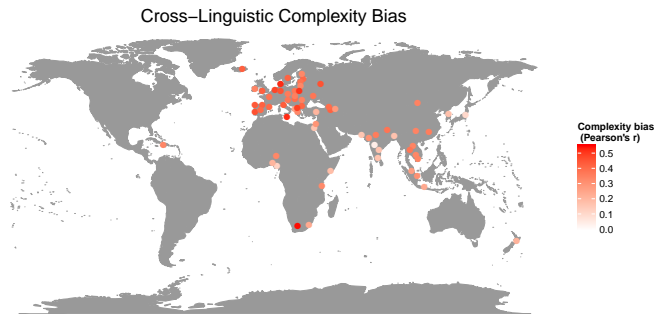


Figure 1  
Magnitude  
of the com-  
plexity bias  
across 79 lan-  
guages. Each  
point corre-  
sponds to a  
language.

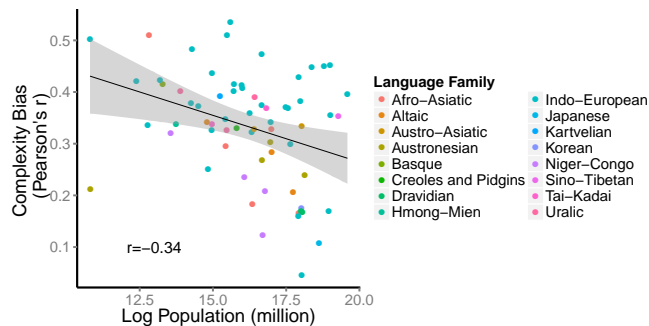


Figure 2  
Relationship be-  
tween complexity  
bias and speaker  
population. Each  
point corresponds  
to a language.

We hypothesized that the same force might influence the degree to which languages encode information density in the lexicon. We calculated the correlation between a language's complexity bias and its population of speakers. Consistent with previous work, we found that languages with more speakers tend to have a smaller complexity bias ( $r = -.34$ ; Fig. 2), and this result remained reliable even after controlling for language family. Thus, learnability pressure may force languages to rely on non-lexical strategies, like speech rate (Pellegrino, Coupé, & Marsico, 2011), to maintain a uniform information density across speech.

## References

- Frank, A., & Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. *Cogsci. Washington, DC: CogSci*.
- Lewis, M., & Frank, M. C. (under review). The length of words reflects their conceptual complexity.
- Lewis, M., Sugarman, E., & Frank, M. C. (2014). The structure of the lexicon reflects principles of communication. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*.
- Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PloS one*, 5(1), e8559.
- Pellegrino, F., Coupé, C., & Marsico, E. (2011). A cross-language perspective on speech information rate. *Language*, 87(3), 539–558.