Frequency, predictability and grammatical asymmetries:
Evidence from Google n-grams

Natalia Levshina

Aims and background

Frequency plays a central role in usage-based linguistics. However, the empirical studies of diverse frequency types and their role in language learning, processing and production are still an emergent field (e.g. Divjak & Gries 2012). The present paper contributes to a more exact understanding of the role of different frequency measures in shaping up language structure. In particular, we focus on the well-known correlation between word frequency and length. Although this correlation has been well-known since Zipf’s seminal work (1949), a more recent study by Piantadosi et al. (2011) has shown that contextual predictability (i.e. the conditional probability of a word given the preceding context) in fact outperforms ‘isolated’ frequency in predicting the word length. More predictable (and therefore less informative) words tend to be shorter, whereas less predictable (and more informative) words are usually longer. These important findings call for a new evaluation of Zipf’s legacy and support the theory of uniform informational density as a means of optimization of human communication (e.g. Jaeger 2010). However, it remains unclear how these findings tie in with a well-known phenomenon in typology, namely, formal asymmetries between members of grammatical categories. For example, singular nouns tend to be formally unmarked and therefore shorter than the corresponding plural forms. This fact has been explained by frequency asymmetries between the forms (e.g. Greenberg 1966; Haspelmath 2008). The present study aims to test whether Piantadosi et al’s results hold both within and across grammatical categories.

Data and methods

We employ the Google books n-grams in English, French, German, Italian, Russian and Spanish (Lin et al. 2012). We compute the normalized and relative (paradigmatically) frequencies, as well as average contextual predictability scores, for samples of wordforms representing different grammatical categories from different word classes. To answer the research question, we use mixed-effects logistic Generalized Additive Models with contrasting category members (e.g. singular or plural) as the response and the frequency and predictability measures as predictors. We also add the diachronic dimension by including the time period as an independent variable.

Preliminary results

Our preliminary results nuance Piantadosi et al’s results in an important way. In particular, we find that formally marked and longer forms of the same words can be in fact more predictable from the context than shorter forms, contrary to what one could expect. In contrast, the frequency of a word form in comparison with its paradigmatic alternative(s) works as predicted by Greenberg and his successors. The results demonstrate that informational density is not the only factor that determines the linguistic form and one should take into account the paradigmatic relationships between forms and categories.
References


