Bringing together corpus linguistics and typology: Frequency, informativity and grammatical asymmetries

Natalia Levshina (Leipzig University, Germany)

Aims and background

Frequency plays a central role in corpus and usage-based linguistics. However, there remain many open questions about the role of diverse frequency types in language learning, processing and production (cf. Divjak & Gries 2012). The aim of the present paper is to contribute 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. This correlation has been well-known since Zipf's seminal work (1935[1968]) as the Law of Abbreviation. It has been shown that the Law of Abbreviation is an absolute language universal (Bentz & Ferrer-i-Cancho 2016).

More recently, a study by Piantadosi et al. (2011) has demonstrated that the average contextual informativity of a word (i.e. the inverse of the conditional probability of the word given the preceding *n*-grams) in fact correlates more strongly with the word length than the context-free probability (i.e. normalized token frequency). 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).

At the same time, it remains unclear how these findings tie in with a well-known phenomenon in typology, namely, formal asymmetries between members of grammatical categories (e.g. Greenberg 1966). For example, it is well known that singular nouns tend to be formally unmarked or less marked than the corresponding plural forms across a variety of languages (e.g. Brunner 2010), e.g. English *chair* – *chairs*, German *Stuhl* – *Stühle*, Russian *stul-Ø* "chair.NOM" – *stul-ja* "chairs.NOM". This fact has been explained by frequency asymmetries between the forms, i.e. by the tendency of the less formally marked forms to have higher relative frequencies than the more marked ones (Haspelmath 2008), in accordance with the principle of economy and minimization of effort. The goal of the present study is to find out whether context-based informativity is more strongly associated with grammatical asymmetries than context-free relative frequency. For this purpose, we perform three case studies, focusing on 1) singular and plural nouns, as in the examples above, 2) positive, comparative and superlative forms of adjectives, e.g. *fine* – *finer* – *finest*, and 3) cardinal and ordinal numerals, e.g. *ten* – *tenth*.

Data and method

In these case studies, we employ the Google Books *n*-grams in English, French, German, Italian, Russian and Spanish (Lin et al. 2012), which are available online at http://storage.googleapis.com/books/ngrams/books/datasetsv2.html (last access 09.01.2017). We select samples of nouns, adjectives and numerals (100-400 words

per class) and compute the relative (paradigmatically) frequencies, as well as average contextual informativity scores, for the singular and plural forms (nouns), different degrees of comparisons (adjectives), and ordinal and cardinal forms (numerals). Lexemes with identical forms, e.g. German *Lehrer* "teacher (SG and PL)", are omitted because of the lack of sufficient morphological information in the data. Following the results presented in Piatandosi et al. (2011), the informativity scores are based on the left-context *n*-grams with $n=3$. To determine which of the two measures is more strongly associated with the formal asymmetries between the less and more marked categories, we use non-parametric paired Wilcoxon tests and mixed-effects binomial and ordinal logistic models. In these models, the contrasting category members (e.g. singular or plural) serve as the response, and the frequency and informativity measures function as predictors. The lexemes (specific nouns, adjectives and numerals) are treated as random effects (intercepts).

Preliminary results

Our preliminary results show that both frequency and informativity behave as expected across the languages: the unmarked and shorter forms are usually more frequent and less informative than the marked and longer ones. Figures 1 and 2, which represent the results for the singular and plural forms of nouns in British English, illustrate the point. Moreover, frequency and informativity are strongly correlated. However, unlike in the study of lexical units in Piatandosi et al. (2012), the more sophisticated informativity measures do not yield a significant improvement in the discrimination between the shorter and longer grammatical categories in comparison with the simpler frequency measures, sometimes performing even worse, as in the illustration. The results thus suggest that informational density is not the only factor that determines the linguistic form and that one should take into account the paradigmatic relationships between forms and categories.

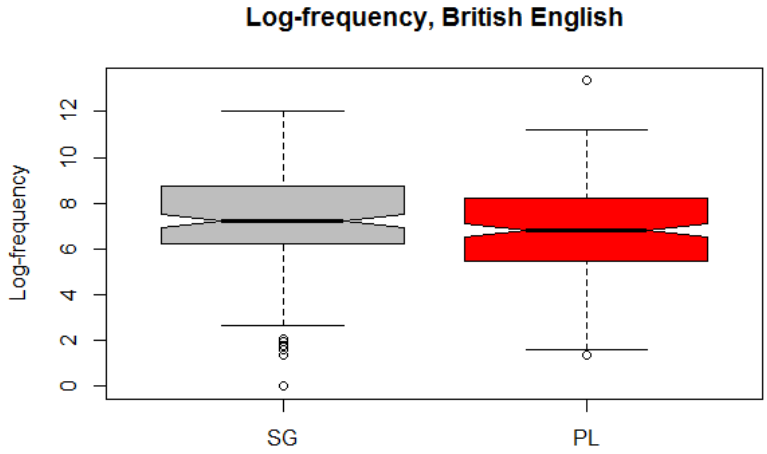

Figure 1. The difference in log frequency between singular and plural nouns in British English. Wilcoxon paired signed rank test: $V = 15436$, $p < 0.001$.

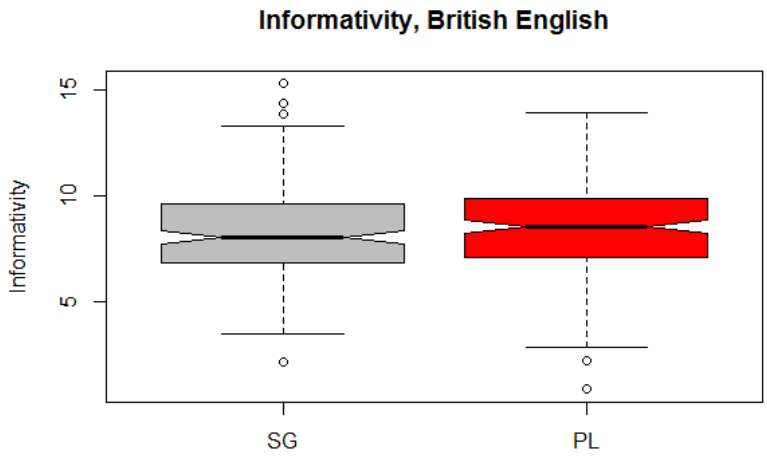

Figure 2. The difference in informativity between singular and plural nouns in British English, based on 3-grams (left context). Wilcoxon paired signed rank test: $V = 8644$, $p < 0.014$.

References

Bentz, Christian & Ferrer-i-Cancho, Ramon (2016). Zipf's law of abbreviation as a language universal. In Bentz, Christian, Jager, Gerhard & Yanovich, Igor (eds.) Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics. University of Tubingen, online publication system. https://publikationen.uni-tuebingen.de/xmlui/handle/10900/68558.

Brunner, J. 2010. Phonological length of number marking morphemes in the framework of typological markedness. In S. Fuchs, P. Hoole, C. Mooshammer & M. Zygis (eds.), *Between the Regular and the Particular in Speech and Language*, 5–28. Berlin: Peter Lang.

Divjak, D. & S. Th. Gries. 2012. Frequency effects in language representation. Berlin: Mouton.

Greenberg, J. 1966. *Language universals, with special reference to feature hierarchies*. The Hague: Mouton.

Haspelmath, M. 2008. Frequencies vs. iconicity in explaining grammatical asymmetries. *Cognitive Linguistics* 19 (1): 1–33.

Jaeger T.F. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology* 61: 23–62.

Lin, Y., J.-B. Michel, E. Lieberman Aiden, J. Orwant, W. Brockman & S. Petrov. 2012. Syntactic Annotations for the Google Books Ngram Corpus. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju, Republic of Korea, 8-14 July 2012, 169–174. Available at http://aclweb.org/anthology/P/P12/P12-3029.pdf

Piantadosi, S., H. Tily & E. Gibson. 2011. Word lengths are optimized for efficient communication. *PNAS* 108(9). Available at www.pnas.org/cgi/doi/10.1073/pnas.1012551108

Zipf G. 1935 [1968]. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Cambridge, MA: MIT Press.