Do morphological oppositions obey Zipf's law of abbreviation? Quantitative evidence from 54 languages

Mittwoch,
04.03.2020
17:30–18:00
ESA1 HG HS M

Aleksandrs Berdicevskis

Språkbanken (The Swedish Language Bank), University of Gothenburg

aleksandrs.berdicevskis@gu.se

Zipf's law of abbreviation (frequent forms are likely to be shorter) is one of the consequences of the principle of least effort (Zipf 1936:30). It is assumed to manifest itself at various language levels (Greenberg 1966; DuBois 1985: 363). It has, for instance, been claimed that more frequent members of morphological oppositions are on average shorter (Haspelmath 2008; Hawkins 2010), e.g. singular forms are usually shorter than plural forms. In this talk, I present the first (to my knowledge) large-scale test of this assumption. Using the Universal Dependencies (v. 2.3) corpora, I test whether the assumption holds across 54 languages and 34 morphological features.

The dataset contains 1110 datapoints (a datapoint in this case is a language-specific feature, e.g. number in Basque, case in German, case in Russian etc.). For every datapoint, I test whether inflectional affixes in the forms that have the most frequent value of the respective feature (e.g. nominative for Russian case) will on average be shorter than those in the forms that have the least frequent value (e.g. dative).

In 64% of datapoints, there is a significant negative correlation between frequency and length, as predicted by the abbreviation law. In 19%, there is a significant positive correlation, in the remaining 17% the correlation is not significant.

Features comply to the abbreviation law to various degree. Aspect, voice, number and degree (of comparison) are most compliant (83–85%), while mood is least compliant (25%, while 30% of datapoints show a significant positive correlation). A closer inspection reveals that one of the most prominent exceptions is the opposition of indicative and imperative: the latter often is shorter, even though it is less frequent than the former.

The Universal Dependencies collection contains several large corpora of ancient Indo-European languages (Latin, Ancient Greek, Old Church Slavonic and Gothic), which enables some (limited) diachronic comparisons. When compared to the modern languages of the respective group (Romance, Greek, Slavic and Germanic), the ancient languages in this particular dataset always show a noticeably smaller compliance to the abbreviation law.

In the talk I also highlight some potential limitations of this corpus-based approach (the Universal Dependencies sample is strongly skewed towards certain groups of Indo-European languages; the corpus annotation is not entirely harmonized across all languages etc.) and discuss how it can be complemented by other approaches (Kanwal et al. 2017).

References: Du Bois, J. (1985). Competing motivations. In J. Haiman (ed.) Iconicity in syntax, 343–365. Greenberg, J. H. (1966). Language universals, with special reference to feature hierarchies. Haspelmath, M. (2008). Creating economical morphosyntactic patterns in language change. In J. Good (ed.) Linguistic universals and language change, 185–214. Hawkins, J. A. (2010). Processing efficiency and complexity in typological patterns. In D. Bakker (ed.) The Oxford handbook of linguistic typology. Kanwal, J., et al. (2017). Zipf's law of abbreviation and the principle of least effort: Language users optimise a miniature lexicon for efficient communication. Cognition 165, 45–52. Zipf, G. K. (1936/2013). The psycho-biology of language: An introduction to dynamic philology.