Predicting native speaker choice: the role of frequency in morpho-syntactic alternations

Abstract:

The issue of how to model native speakers’ preferences for one or the other alternant in a given context has received a substantial amount of attention (Bresnan et al. 2007, Bresnan & Ford 2010, Divjak et al. 2016). Multivariate corpus-based and experimental research has revealed a number of variables that significantly affect subjects’ preferences across a range of different languages. However, there is no clear agreement as to which out of a number of competing frequency metrics is best suited to predict native speaker behaviour. We aim to fill this gap by looking at a morpho-syntactic alternation between the adessive case and the postposition peal ‘on’ in Estonian. The frequency metrics under consideration fall into two broad groups: (1) collocation and collostructional metrics (Gries & Stefanowitsch 2004, Schmid & Küchenhoff 2013, Gries & Ellis 2015), (2) information-theoretic metrics like entropy and surprisal (Hale 2016). Based on prior research, we expect language users to show a dispreference for constructions which are more ‘surprising’ in an information-theoretic sense (Smith & Levy 2013).

In order to address the role of frequency in predicting native speaker choice, we present the results of a forced choice task carried out with 103 native speakers of Estonian. The participants were presented with 60 attested sentences in which the original construction was replaced with a blank. There were 30 experimental items and 30 filler items in the experiment. The experimental items were randomly sampled from five equal probability bins defined by a logistic regression model fitted to the corpus sample. For each sentence an alternative paraphrase was constructed for the original construction and both alternatives were presented together with the original sentence context. Participants were asked to choose which of the two constructions suits into the blank better. The experiment was designed and distributed using the online platform qlaara (https://qlaara.com/).

The experimental data were analysed using mixed-effects logistic regression. All of the frequency metrics of interest were extracted from the Balanced Corpus of Estonian (BCE 2015; size 15 million words in total) and etTenTen (270 million words from 686,000 webpages in Estonian). The results suggest that the native speakers of Estonian are attuned to the global frequency of the adessive construction, which is 10 times more frequent than the peal construction in the locative function. Moreover, language users’ preferences are influenced by the relative frequencies with which certain nouns appear with different locative cases and postpositions.
References:


