A multivariate analysis of the alternation between the adessive case and postposition *peal* ‘on’ in Estonian dialectal data

The present paper hopes to make a methodologically sound contribution to the study of constructional alternations in the non-standard varieties in the East of the Circum-Baltic area by taking a usage-based and variationist perspective. The paper takes a look at a constructional alternation in Estonian dialectal data using the corpus of Estonian dialects (CED 2015). The focus is on the parallel use of the adessive case construction (example 1) and the adposition *peal* ‘on’ construction (example 2):

1. **suured**  _liha_  _kaosid_  _ollid_  _lava_  _pial_ (MUH)
   big-NOM meat-GEN bowl-NOM be.PST table-GEN on
   ‘the big bowls of meat were on the table’

2. **nied**  _ölid_  _kaa_  _lauwal_  _sis_ (LÜG)
   they-NOM be.PST also table-ADE then
   ‘these were also on the table then’

The data sample contains a total of 2,131 occurrences of the adessive case and the adposition *peal* ‘on’ across 286 speakers and 10 dialectal areas. A previous study on the same dataset (Klavan et al. 2015) showed that the alternation between the synthetic adessive construction and the analytic *peal* ‘on’ construction is not free. The study identified a number of semantic and morpho-syntactic factors that influence the choice between the two constructions in non-standard, spoken language confirming the claim made in numerous other studies that the choice between alternative constructions is driven by a multitude of factors (e.g. Divjak & Arpe 2013, Bresnan 2007, Bresnan et al. 2007, Bresnan & Ford 2010, Szmrecsanyi 2013). The data in Klavan et al. (2015) were analysed using mixed-effects logistic regression (Harrell 2001; Pinheiro & Bates 2002; Hosmer et al. 2013) and the ‘tree & forest’ method (Breiman 2001; Strobl et al. 2009). The models fitted to the data were able to predict the choice between alternative constructions with a 94% of classification accuracy. Klavan et al. (2015) showed that specific lemma (and to a lesser extent the specific informant) contribute significantly to predicting the choice between the two alternatives. In addition, the models also confirmed the importance of the linguistic fixed-effects. The variables of length, complexity, and type of Landmark as well as verb group and dialect all play a role in the variation between the adessive and *peal* ‘on’.

Klavan et al. (2015) conclude that the mixed-effects logistic regression model as well as the tree-based models reported in their paper have an excellent fit and a very high prediction accuracy. Still, there are a number of potentially relevant predictors that are missing from their models, but which are shown to make significant contributions to models fitted to other alternation phenomena. Most importantly, the present paper looks at frequency and priming effects. For example, Szmrecsanyi (2005, 2006), Hinrichs & Szmrecsanyi (2007), Bresnan & Ford (2010: 174) show that structural parallelism or persistence is an important predictor in syntactic choice. According to Szmrecsanyi (2005: 113), speakers re-use a recently used or heard linguistic construction whenever they can. Furthermore, a well-known fact about language use concerns human sensitive to frequency information (Ellis 2002; Divjak & Gries 2012; Gries & Divjak 2012; Divjak & Caldwell-Harris 2015). The aim of the present paper is, therefore, to take this line of analysis further by including frequency and persistence in the mixed-effects logistic regression model and the random forest model fitted by Klavan et al. (2015).
References
Strobl, Carolin; Boulesteix, Anne-Laure; Zeileis, Achim & Hothorn, Torsten. 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC bioinformatics 8 (1).