ARE CORPUS-BASED PREDICTIONS MIRRORED IN THE PREFERENTIAL CHOICES AND RATINGS OF NATIVE SPEAKERS? PREDICTING THE ALTERNATION BETWEEN THE ESTONIAN ADESSIVE CASE AND THE ADPOSITION *PEAL* 'ON'

Jane Klavan
University of Tartu

Ann Veismann
University of Tartu

Abstract. Recent work in usage-based linguistics stresses the importance of combining corpus-based analyses with experimental studies. A number of studies have compared the performance of a corpus-based statistical model against the behaviour of native speakers in a linguistic experiment. The present paper takes this line of analysis further by combining corpus-based work with two sources of experimental data. A mixed-effects logistic regression model is fitted to the corpus data of the Estonian adessive case and the adposition *peal* 'on' in present-day written Estonian. In order to evaluate the goodness of the corpus-based model, its performance is compared to the behaviour of native speakers in a forced choice task and a rating task.

Keywords: alternating constructions, corpus linguistics, forced choice task, rating task, statistical modelling, Estonian

DOI: https://doi.org/10.12697/jeful.2017.8.2.03

1. Introduction

Recent years have witnessed a renewed focus on methodological pluralism within the framework of usage-based linguistics. The "quantitative turn" (Janda 2013) has brought about an impressive growth in the number of published studies that combine corpus-based data with behavioral data. One prolific area pertains to the discussion of how corpus-based frequency estimates relate to experimental findings, especially acceptability judgements (see Divjak 2016 for a recent overview). There has also been an exponential growth in published studies that use probabilistic statistical classification models to analyse linguistic data; see Klavan and Divjak (2016) for an overview. Still, only a small number of these studies have compared their findings with behavioral data (Roland *et al.* 2006, Wasow and Arnold 2003). Few have used authentic corpus data for this purpose (Arppe and Järvikivi 2007, Bermel and Knittl 2012, Divjak and Gries 2008) and even fewer have directly compared the performance of a complex corpus-based model against humans using authentic corpus sentences (Arppe and Abdulrahim 2013, Bresnan 2007, Divjak *et al.* 2016). In addition to statistical modelling, there are other ways how corpus and experimental data can successfully be combined. Gilquin and Gries (2009) provide at the time a comprehensive overview of these studies, both from the perspective of corpus linguistics as well as psycholinguistics. Whether we start off from a corpus analysis or a psycholinguistic experiment, we align with those who argue that the study should be supplemented with behavioral data or data on actual language use respectively.

The few previous studies that have pitted corpus-based models against behavioral data using authentic corpus sentences have shown that an adequately constructed probabilistic model based on an extensively annotated corpus data can perform at a more or less equal level to human beings (Arppe and Abdulrahim 2013, Bresnan 2007, Bresnan and Ford 2010, Divjak *et al.* 2016). These studies show that the probability estimates calculated by the corpus-based models are mirrored in the ratings and the proportion of choices made by the native speakers. Furthermore, Bresnan (2007) and Arppe and Abdulrahim (2013) have shown that the responses given by the native speakers can be successfully explained as a function of the original corpus model predictors. Thus far, experimental "validation" of corpus-based models has been restricted to off-line tasks (forced choice task and acceptability ratings), although on-line studies are also being run (Arppe *et al.* 2012, Bresnan and Ford 2010, Ford and Bresnan 2013). The main difference lies in what can be tested with each type of method. Off-line studies concentrate on the full set of predictors of the corpus-based model, while on-line studies – for various (practical) reasons – pick out a few predictors from the full model for the experiment. The present study also uses off-line tasks, because we are concerned with comparing the performance of the entire model vis-à-vis the performance of native speakers. Nevertheless, we acknowledge the need for future research to employ more advanced experimental techniques to validate the corpus-based model presented in this paper.

Our paper makes three important contributions to the discussion of comparing or contrasting corpus-based evidence against behavioral evidence: 1) it explicitly pitches the performance of a complex corpus-based model against human behaviour using authentic corpus sentences; 2) it looks at behavioural data from two experiments (a forced choice task and a rating task); 3) it uses data from a less widely studied language. The third aspect – looking at a language other than English – provides important typological evidence and gives us confidence that the findings apply cross-linguistically. The first two aspects enable us to address the critical question whether the model we have fitted is cognitively plausible. When we talk about the cognitive plausibility of the model, we mean here, and elsewhere in this paper, the end-result of the modelling process rather than the modelling process itself. Thus, we are concerned with assessing whether the corpus-based model is sufficiently accurate and efficient when used for classification and prediction, whether the features picked up by the model help us to capture what actually goes on in language. We argue, similarly to others (e.g. Klavan and Divjak 2016, Divjak *et al.* 2016), that without behavioral data it would be very difficult if not impossible to provide an adequate assessment of a corpus-based model. Linguistic experiments are necessary to calibrate our models – sometimes models are very accurate, and sometimes they appear to be less accurate; in order to set "upper and lower boundaries to what could be psychologically relevant" we need behavioral data to evaluate the corpus-based model (Klavan and Divjak 2016).

The aim of our paper is to evaluate the performance of a corpus-based mixed-effects logistic regression model by comparing the corpus-based predictions against the preferential choices and ratings of native speakers in linguistic experiments. It is assumed that the predictions made by the corpus-based model are mirrored in the behaviour of native speakers. More specifically, we are interested in finding out whether the corpus-based predictions are reflected equally well in native speakers' preferential choices and their ratings. If not, we are interested in determining the place and source for divergence. Ultimately, we seek to establish which if either source of behavioral data provides a good reflection of the corpus-based data. To this end, we present a multivariate statistical analysis of an extensively annotated dataset from present-day written Estonian on the parallel use of the adessive construction and the postpositional *peal* 'on' constructions. Both constructions can be used to express the location of one object on top of another object. We refer to the two constructions as alternating constructions.

The rest of the paper is structured as follows: In Section 2 we briefly introduce the data, the annotation schema and the mixed-effects logistic regression model fitted to the corpus data. In Section 3 and Section 4 we provide experimental evidence from two linguistic experiments. We emphasise three important findings: 1) forced choice data compared to rating data provides a slightly better reflection of the corpus-based model; 2) there is a strong positive correlation between the choices made and the ratings given by the native speakers of Estonian; and 3) out of the two alternative constructions, the adessive construction is the default choice for native speakers. These issues are elaborated in Section 5 where we discuss the findings of our study. The paper ends with a conclusion.

2. The Estonian adessive and *peal* construction: a corpus-based model

Examples (1) and (2) illustrate the alternation between the Estonian adessive case construction and the postpositional construction with *peal* 'on'. Both constructions express a situation where an object (the Trajector[1], henceforth TR) is located on top of another object (the Landmark, henceforth LM).

(1) *Raamat* *on* ***laua-l***.
book.SG.NOM be.PRS.3SG table-SG.ADE
'The book is **on the table**.'

(2) *Raamat* *on* ***laua*** ***peal***.
book.SG.NOM be.PRS.3SG table.SG.GEN on
'The book is **on the table**.'

It has been claimed in Estonian reference grammars that the meaning of adpositions is more concrete and specific than that of the cases, while the usage range of the latter is much broader (Erelt *et al.* 1995: 33–34, Erelt *et al.* 2007: 191). This is in line with the general claims made in the literature concerning the differences between adpositions and case affixes (Comrie 1986, Hagège 2010, Lestrade 2010). In other Finno-

[1] Langacker's (2008: 70) terminology is used to refer to the two most fundamental notions in relational expression: Trajector and Landmark. Trajector is the entity whose location or motion is of relevance; Landmark is the reference entity in relation to which the location or the motion of the Trajector is specified.

Ugric languages, Bartens (1978) and Ojutkangas (2008) have found that the analytic adpositional construction, compared to the synthetic case construction, places more stress on location and is used together with smaller, manipulable things as Landmarks.

Previous studies for the alternation between the Estonian adessive and *peal* have shown that the probability of using one or the other of the two alternating constructions is associated with a number of semantic and morpho-syntactic properties. Klavan (2012) reports a simple binary logistic regression model with 6 predictors fitted to present-day written Estonian: mobility of Landmark, verb group, length of the Landmark phrase in syllables (logarithmically transformed), morphological complexity of Landmark, word class of Trajector, and the relative position between Trajector and Landmark. The forced choice task and the acceptability rating studies reported in Klavan (2012) confirm the relevance of the mobility or type of the Landmark, but not the length of the Landmark phrase. A further corpus study by Klavan *et al.* (2015) for non-standard spoken language, reports multivariate models that can predict the choice between the two alternative constructions with a 94% classification accuracy. Although it was shown that the specific lemma contributes significantly to model fit, mixed-effects logistic regression also confirmed the importance of the linguistic fixed-effects. More specifically, Klavan *et al.* (2015) show that length, complexity, type of Landmark, verb group and dialect all play a role in the variation between the adessive and *peal*.

2.1. The corpus-based mixed-effects logistic regression model

Building on earlier work (Klavan 2012, Klavan *et al.* 2015), we show that it is possible to predict which of the two constructions is used in written present-day Estonian with 80% accuracy. We used mixed-effects logistic regression (Baayen *et al.* 2008, Pinheiro and Bates 2000) to analyse the corpus data from Klavan (2012). Klavan (2012) extracted contextual data, i.e. semantic and morphosyntactic information found within clause boundaries, for both the adessive and *peal* constructions. The data come from the fiction and newspaper subcorpora of the Morphologically Disambiguated Corpus (MDCE 2015; size 215,000 words) and the Balanced Corpus of Estonian (BCE 2015; size 10 million words). The two corpora contain literary (108 authors) and newspaper texts published from 1980s to 2000s. A random sample of 450 extractions per construction was selected for the multivariate analysis. The 900 tokens were manually tagged for 11 semantic and 10 morphosyntactic variables or features which capture the information provided at the clause level concerning the properties of the Trajector and Landmark phrase, type of verb and type of clause. There were a total of 21 nominal variables amounting to 47 distinct variable categories or contextual properties (see Klavan 2012: 70–92 for details).

Model building is a crucial step in logistic regression analysis, but opinions are divided as to which strategy is the best one for arriving at an optimal model (Burnham and Anderson 2002, Harrell 2001, Hosmer *et al.* 2013). In this paper, we have adopted a stepwise model simplification strategy, where the minimal adequate model is selected from a large set of more complex models (Crawley 2007: 323–386). The stepwise progression from the maximal model including all of the 21 variables and their interactions to the minimal adequate model was made on the basis of deletion tests (*F*-tests or chi-square tests). An explanatory variable was retained in the model only if it significantly improved the fit of the model. Any redundant parameters (non-significant interaction terms and non-significant explanatory variables) were removed. Altogether 4 variable categories (1 semantic and 3 morphosyntactic) and LEMMA as a random effect were retained in the final, minimally adequate model fitted to the corpus data. An overview of the predictors retained in the final model and their levels are given in Table 1.

Table 1. Overview of the predictors and their levels used in the corpus-based mixed-model.

Predictor	**Levels**
CONSTRUCTION (dependent variable)	adessive, *peal*
LM_LENGTH (length of the Landmark phrase in syllables; logarithmically transformed)	ranging from 1 to 41 syllables
LM_COMPLEXITY (morphological complexity of the word used in the adessive or *peal* construction)	compound, simple
LM_MOBILITY (mobility of Landmark)	mobile, static
TR_WORDCLASS (word class of the Trajector phrase)	noun, other
LM_LEMMA (lemma of the word used in the adessive or *peal* construction)	397 lemmas

The optimal logistic mixed-model for the adessive and *peal* alternation is described by the following formula:

CONSTRUCTION ~ LM_LENGTH + LM_COMPLEXITY + LM_MOBILITY + TR_WORDCLASS + (1|LM_LEMMA)

Model accuracy is evaluated by two measures – percentage of overall accuracy and the C measure (the index of concordance or the area under the receiver operating characteristic curve; Hosmer *et al.* 2013: 173–182). Overall accuracy is estimated by cross-tabulating the two possible outcomes by high and low probabilities based on a cut-off point set at 0.5. The model makes a correct prediction if the estimated probability for *peal* construction is greater than or equal to 0.5 and the *peal* construction was actually observed in the data. The overall accuracy or model fit (since the model is trained and tested on the same 900 instances) of the corpus-based mixed-model is 80% and the *C* measure is 0.88. Both measures indicate that the model is doing a good job as a classifier – a *C* value between 0.7 and 0.8 is generally considered as acceptable discrimination, while a value above 0.8 is deemed as excellent discrimination (Hosmer *et al.* 2013: 177). The improvement rate of the model (calculated by dividing accuracy by baseline, which for the present dataset is 50%) is 1.6. It seems reasonable to assume that the model provides a reasonable fit. There is a 30% increase in classification accuracy compared to random classification, which is a very good result considering that we are predicting a choice between two near-synonymous constructions, i.e. relatively similar underlying probabilities are only to be expected since, in principle, both alternatives can be used in all of the studied contexts.

2.2. Variable importance in the corpus-based mixed-effects logistic regression model

This section takes up the question of the interpretability of the model and looks at the contribution to model fit by individual linguistic predictors. Decrease in the Akaike Information Criterion (AIC; Hosmer *et al.* 2013: 120) is taken as an indicator of the importance of a particular predictor (Baayen *et al.* 2013: 264). AIC is used to compare the fit of models with different number of parameters – a smaller value is taken as an indication of a better model fit. Individual parameter estimates were tested by the likelihood ratio test, a test based on the difference in deviances; the results are given in Table 2. The first column in Table 2 shows the order in which the predictors were added to the intercept only model (the null model). The last column lists the reduction in AIC – the larger the reduction in AIC once a specific predictor is added, the more important the predictor is.

Table 2. Model comparison statistics for the mixed-model of corpus data.

	logLik	Chisq	Chi.Df	*p*-value	Reduction in AIC
LM_LEMMA	–589.97				65.7
LENGTH	–553.29	73.356	1	0.000	71.4
COMPLEXITY	–534.22	38.154	1	0.000	36.2
MOBILITY	–524.86	18.717	1	0.000	16.7
TR_WORDCLASS	–517.00	15.716	1	0.000	13.7

Somewhat surprisingly, we can see from Table 2 that the decrease in AIC for the fixed-effect of LENGTH (71.4) is larger than the decrease in AIC for the random effect of LEMMA (65.7). The contribution made by other predictors to model fit is considerably lower: COMPLEXITY (36.2), MOBILITY (16.7), TRWORDCLASS2 (13.7). The Shapiro-Wilk test indicates that the normality assumption of the random lemma-specific intercepts is slightly violated (W = 0.98894, *p*-value = 0.0189). However, the misspecification of the distribution of random effects has little effect on estimates of covariate effects (McCulloch and Neuhaus 2011; Neuhaus *et al.* 2013). It is therefore concluded that the slight violation of the normality assumption of the random lemma-specific intercepts does not pose problems for interpreting the main effects of the model.

As to the specific predictions made we may inspect the estimated coefficients and the main effects for the mixed-model of the corpus data. The coefficients in Table 3 indicate that the adessive construction is preferred when the Landmark is a compound denoting a static place. Conversely, the *peal* construction is preferred when the Landmark is a mobile thing denoted by a short, simple word. Nominal Trajectors predict the adessive, while pronominal and verbal Trajectors predict the *peal* construction.

Although mathematically and statistically speaking we may conclude that we were able to fit a "good" model, one pertinent question can be raised at this point: is human performance comparable to the corpus-based model? As linguists we are interested in finding a model that is sufficiently accurate while at the same time giving us cognitively plausible information about the linguistic phenomenon.

Table 3. Coefficients for re-calibrated logistic regression model of corpus data.

	Estimate	Std. Error	z-value	*p*-value
Intercept	0.244	0.387	0.630	0.5288
LENGTH	–1.075	0.179	–5.991	0.0000
COMPLEXITY = simple	1.517	0.300	5.052	0.0000
MOBILITY = static	–0.958	0.219	–4.363	0.0000
TR_WORDCLASS = other	0.730	0.189	3.858	0.0001

We now turn to linguistic experiments as one possible source for finding answers to this question. The specific question we are interested in is if and how well the corpus-based predictions are mirrored in the forced choice task and the rating task. As a further aim, we also look at the convergence or divergence between the two sources of experimental data. We hypothesise that the Estonian speakers implicitly know the usage patterns of the two constructions and use this knowledge to predict syntactic choices and rate the alternative constructions. In other words, we expect there to be an alignment between probabilities of the corpus-model and the choices and ratings of the native speakers of Estonian.

3. Design of the experiments

3.1. Forced choice task

Materials. The experiment consisted of 30 corpus sentences with a blank for the original construction followed by the two constructional alternatives (see Appendix 1A for an example). The sentences (i.e. experimental items) were randomly sampled from five equal probability bins (6 sentences per each probability bin) defined by the mixed-effects logistic regression model in the 900-observation dataset. The sampled stimuli therefore represent the full probability scale and ranged from sentences where one construction was very probable (near-categorical preferences) to sentences where both constructions were equally probable (approximately equal probability estimates for both choices). The item probabilities estimated by the mixed-effects corpus model are given in Figure 1 together with the lemma of the Landmark phrase (LM_LEMMA). The higher the probability on the y-axis, the more probable it is that the *peal* construction is chosen. We will henceforth use word lemmas as shorthand for the experimental stimuli. However, it should be kept in mind that the lemma refers to the entire sentence used as stimuli in the forced choice task. The full list of experimental items together with the sentence context, the lemma and the English translation is given in Appendix 2.

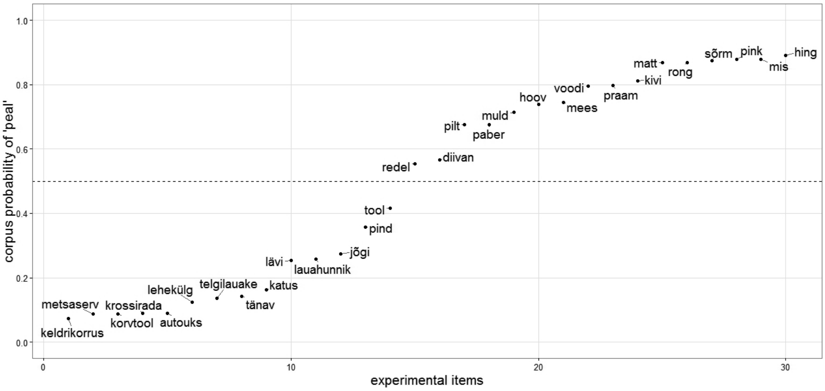


Figure 1. Estimated probabilities for experimental items.

Since the stimuli were randomly sampled, the two constructions are not represented in equal numbers: there were 16 sentences with *peal* construction and 14 with adessive. For each sentence an alternative paraphrase was constructed for the original construction and both alternatives were presented together with the original sentence context. Items were pseudo-randomized so that no two sentences from the same probability bin followed each other. There were four versions of the questionnaire to diminish potential order effects. The order of construction choices was alternated between the versions; for example, in versions 1 and 3 the adpositional construction was given first for item$_a$, in versions 2 and 4 the adessive construction was presented first for the same item. There were no control or filler items.

Participants. 96 native speakers of Estonian were recruited via the Internet using social media (Facebook, mailing list services). They were randomly assigned to one of the four versions of the experiment (v1 = 22, v2 = 29, v3 = 24, v4 = 21). The participants (47 males, 49 females) ranged in age from 18 to 54 (mean 29, SD = 9.5).

Procedure. Participants were asked to choose which of the two constructions suits into the blank better. The two constructional alternatives were presented next to each other horizontally after the corpus sentence with a blank (see Appendix 1A). Participants saw only one sentence at a time and were not able to go back and change their answers. Each subject completed a questionnaire with the same 30 sentences. The questionnaire was designed and distributed using the online survey service PsychData (https://www.psychdata.com). On average, it took about 10 minutes for participants to complete it.

3.2. Acceptability rating task

Materials. The experimental items used in the acceptability rating task were the same as the 30 items used in the forced choice task (Figure 1). For each of the original experimental item an alternative paraphrase was constructed. The adessive and *peal* constructions were separated from the rest of the sentence by square brackets (see Appendices 1B and 1C for examples of the materials). It was decided not to show both alternatives to one and the same participant. The 60 experimental items were divided into two lists of 30 items each, so that each experimental item appeared only once per list. The order of the items was randomized between the four versions of the two lists. There were eight lists all together. There were no control or filler items.

Participants. 98 native speakers of Estonian were recruited via the Internet using social media (Facebook, mailing list services). They were randomly assigned to one of the eight lists of the experiment: version 1a = 12, version 1b = 13, version 2a = 13, version 2b = 12, version 3a = 12, version 3b = 12, version 4a = 12, version 4b = 12. The participants (50 females, 48 males) ranged in age from 15 to 66 (mean 31, SD = 10.7).

Procedure. Participants were instructed to rate the naturalness of the phrase between the square brackets on a 10-point scale ranging from *väga kummaline* 'very strange' (corresponding to 1, on the extreme left) to *täiesti loomulik* 'completely natural' (corresponding to 10, on the extreme right). It was explicitly stated in the instructions of the experiment that the focus of the study is on the alternation between the adessive and *peal* constructions. Participants saw only one sentence at a time and were instructed not to go back and change their answers. Before the actual experiment itself, the participants had four practice trials. The results of the practice trials were not taken into account in the subsequent data analysis. The questionnaire was designed and distributed using the online survey service Google Forms (https://www.google.com/forms/about/). On average, it took about 10 minutes for participants to complete it.

4. Results: pitting corpus-based predictions against choices and ratings

When analysing the results of the two experiments, we make use of exploratory data analysis techniques. Our main aim is to assess how well are the corpus-based predictions mirrored in the forced choice data and the rating data (Section 4.1). Part of the data analysis is also devoted to comparing forced choice data against the rating data (Section 4.2), but we only focus on the main trends. A detailed analysis of how and why the two sources of data converge or diverge merits a separate study – something which we leave for the future. We also look at whether the four predictors picked up by the corpus-based model exhibit similar trends in the forced choice data and rating data (Section 4.3). Again, due to the limits on space we only use exploratory techniques and leave a more complex analysis employing mixed-effects regression, repeated measures ANOVA, Naive Discriminative Learning and other relevant techniques for the future.

Before turning to the analysis of the results, a note on the rating data is in order. Following Divjak *et al.* (2016: 25), the raw acceptability ratings were residualised against participant and position of the experimental items in the experiment: first, ratings were regressed on participant and position and the residuals from this regression analysis were then used in subsequent data analysis. This way we may be confident that each rating is "free of differences in how participants used the scale, or how their ratings changed over the course of the experiment" (Divjak *et al.* 2016: 25). Following Divjak *et al.* (2016: 25), we also rescaled the residualised ratings so that each participant used the entire scale (1–10).

4.1. Agreement between corpus model predictions and native speaker behaviour

In order to compare the agreement between corpus predictions, on the one hand, and native speaker choices and ratings, on the other hand, graphical explorations of the experimental data are presented. One possible way to explore the experimental data is to look at the log of the ratio of adessive choices or ratings and *peal* choices or ratings pitted against the probabilities of the mixed-model fitted to corpus data.[2] This line of analysis allows us to assess the agreement between the predictions made by the corpus-based models against native speaker choices and ratings for the set of 30 experimental items. Figure 2 displays the probabilities of the *peal*-construction as estimated by the corpus-based model for each of the 30 items plotted against the log odds of adessive vs *peal* from the forced choice task (the first plot in Figure 2) and the log odds of adessive vs *peal* from the rating task (the second plot in Figure 2). We use the LM_LEMMA as a shorthand for the experimental items (see Appendix 2 for the sentence contexts and English translations).

The two plots in Figure 2 indicate that the corpus model and native speakers are largely in agreement as to which construction is the preferred one for the 30 experimental items: there is a strong association between the corpus probabilities and the log odds of choices ($r = –0.79$, $p = 0$) and ratings ($r = –0.73$, $p = 0$). The negative sign of the correlation indicates that the lower the probability of the *peal* construction (the vertical dimension or y-axis in Figure 2), the more likely it is that the adessive construction was chosen or received a higher rating.

The results of both tasks suggest that the default choice for native speakers is the adessive construction. This can be seen from both plots in Figure 2, since the majority of the dots fall on the right side of zero on the x-axes. These data points have positive log odds or are very close to 0 indicating thus the predominance of the adessive construction. Participants frequently chose the adessive construction or rated it higher for items where both the original as well as the predicted construction was the *peal* construction. One possible explanation for this result is the fact that the adessive construction is around 10 times more frequent than the *peal* construction in present-day written Estonian (Klavan 2012: 182–183). It is clear that native speakers are attuned to such global frequency information, the corpus-based model, however, was fed an equal number of both constructions; we will return to this issue in the discussion. We will now take a closer look at how much the choices and ratings converge across the experimental items.

[2] The log odds are calculated by hand. In the forced choice data, 1 is added to all counts before taking the log in order to avoid dividing by zero; it is calculated as follows: logit = log((number of adessive constructions + 1)/(number of *peal*-constructions + 1)). For the rating data, the log odds ratio is calculated as follows: logit = log(mean residualised rating for adessive construction/mean residualised rating for *peal* construction).

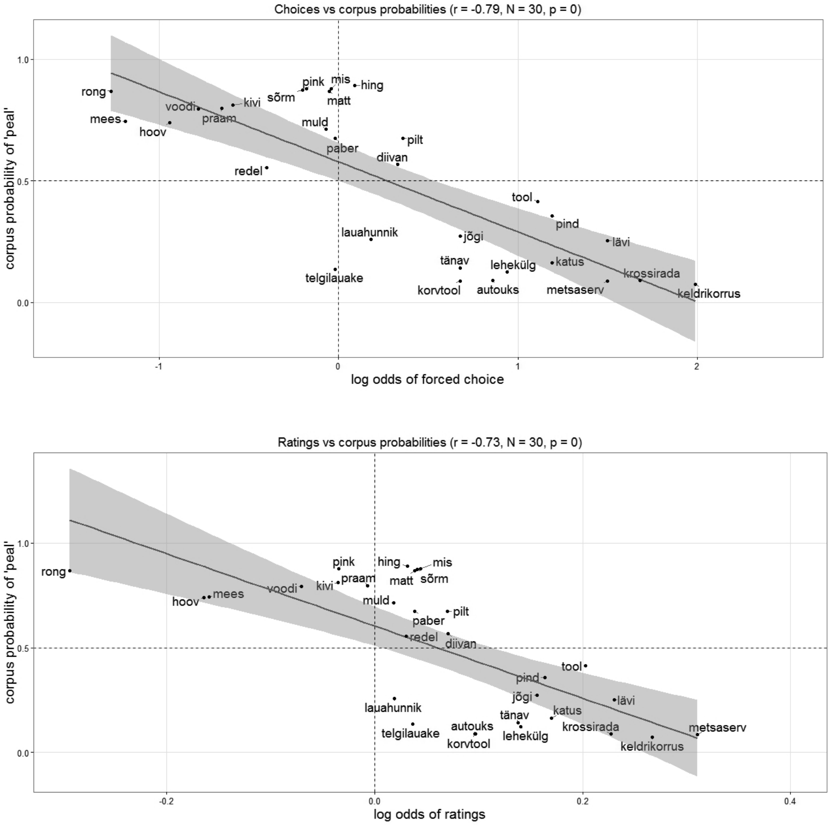


Figure 2. The log odds (of adessive vs *peal*) for each of the 30 experimental items plotted against the respective corpus probabilities of the *peal* construction estimated by the corpus model and the pairwise Pearson correlations. The cut-off point for the horizontal dimension is zero: a dot that falls to the right of zero indicates either that the proportion of adessive choices is higher than the proportion of *peal* choices (in case of the forced choice data on the first plot) or that the adessive construction has received a higher (residualised) rating compared to the *peal* construction (the rating data on the second plot), whereas a dot to the left of zero indicates the predominance of the *peal* construction. The vertical dimension is centred at 0.50 – for dots that are below 0.50 the respective corpus model predicts adessive construction and for dots above 0.50 the *peal* construction.

4.2. Agreement between choices and ratings

In addition to comparing experimental data against corpus data, we may also explore the experimental data by comparing the two types of experimental data against each other. To recap, the experimental items were the same in both tasks. We may first take a look at a similar plot as those presented above. Here, in Figure 3 the log odds of the residualised ratings (y-axis in Figure 3) are plotted against the log odds of the choices (x-axis in Figure 3) for the 30 experimental items (we use lemmas as shorthand for the entire sentence used as stimuli in the two experiments; see Appendix 2 for details and for English translations).

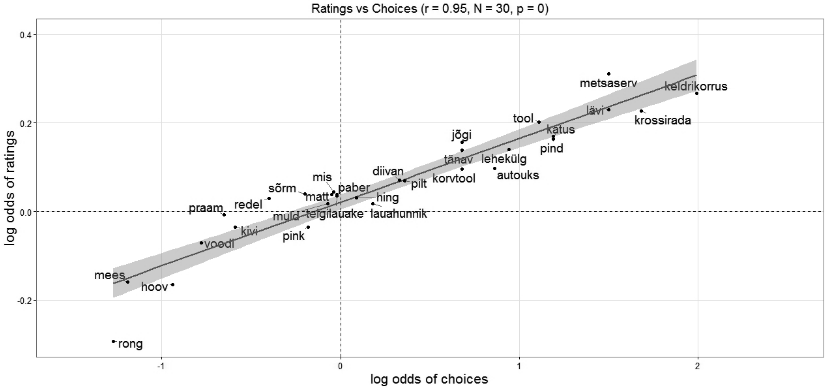


Figure 3. The log odds (of adessive vs *peal*) for each of the 30 experimental items and the pairwise Pearson correlation between residualised ratings and choices. The cut-off point for both the horizontal and vertical dimension is zero: a dot that falls to the right of or above zero indicates the predominance of the adessive construction, whereas a dot to the left of or below zero indicates the predominance of the *peal* construction.

It is seen from Figure 3 that the choices made and the ratings given by the native speakers agree very strongly across the 30 experimental items ($r = 0.95$, p = 0). The plot can be read by cutting it up into four sections: (i) in the lower left grid (to the left of and below zero) we find items that are preferred in the *peal* construction in both tasks; (ii) in the upper left grid (left of and above zero) are items that are preferred in the *peal* construction only in the forced choice task – in the acceptability task these items have received a higher rating with the adessive construction; (iii) in the upper right grid (to the right of and above zero) we find items that are preferred in the adessive construction in both tasks; (iv) there are no items in the lower right grid (to the right of and below zero) – in other words there were no items that preferred the adessive construction in the forced choice task but not in the acceptability task. The fact that majority of the items fall either in the upper right or lower left grid on Figure 3 shows that the forced choices converge with the ratings. In general, we are seeing what we already saw above – the overall preference for the adessive construction. We will come back to this issue in our discussion (Section 5) and presently move on with the presentation of the results by looking in greater detail at the convergence and divergence between choices and ratings at the level of the specific items.

Figure 4 displays two plots which dismantle to some extent the two sources of data – it is clear that the difference between the choices and ratings is more pronounced for some items and less distinctive for other items. The first plot in Figure 4 is a barplot that represents the count of the two constructional choices (for each bar, N = 96) across the 30 items; the second plot in Figure 4 represents the mean residualised ratings for the two constructions across the 30 items. The items on the x-axis on both plots in Figure 4 are given in the order of corpus-based probability for the *peal* constructions – the items on the far left are items for which the corpus model predicted a low probability of the *peal* construction and items on the right for which a high corpus-based probability was assigned.

From the barplot in Figure 4 we can deduct that there is a noticeable trend as we move from the left of the x-axis to the right – the height of the dark grey bars (the number of times the adessive construction was chosen) diminishes as the corpus-based probability estimates of the *peal* construction increase. This result echoes the first plot in Figure 2 – there is a significant correlation between choices and corpus-based probability estimates. For the second graph in Figure 4, this trend is not as noticeable – it exists for the *peal* construction, i.e. the filled triangles appear lower on the y-axis on the left end of the x-axis and higher on the right end of the x-axis, but the position of the filled circles (the residualised mean ratings for the adessive construction) remains relatively high across the entire graph (except for *hoov* 'yard', *mees* 'man', *voodi* 'bed', and *rong* 'train'). This result is also mirrored in the second graph of Figure 2 – there is a correlation between ratings and corpus-based probability estimates, but it is less strong than the correlation between choices and probabilities.

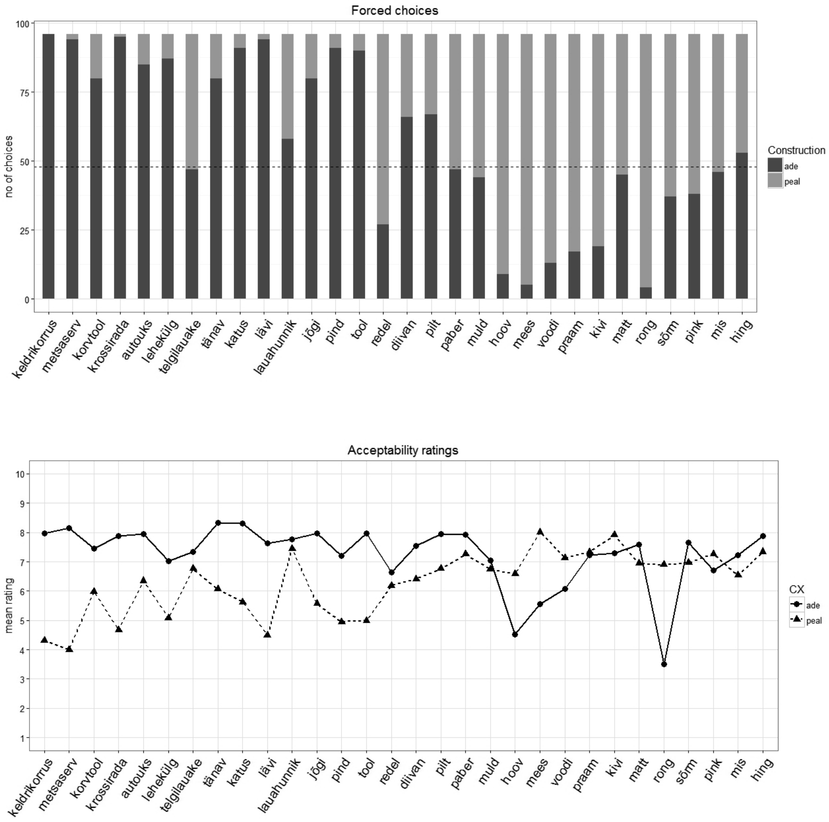


Figure 4. The proportion of choices and the residualised mean ratings plotted against the 30 experimental items according to the two constructions (dark grey or filled circle – adessive, light grey or filled triangle – *peal*); the experimental items are given in the order of corpus-based probability estimates.

Let us first look at the items where the two sources of data – choices and ratings – converge the most. We may look at two distinct groups here: a) items for which both sources of data show that there is a clear preference for one of the two constructions, and b) items for which there is no clear difference between the choices and ratings. From the first subgroup, three items show a pronounced preference for the *peal* constructions: *rong* ‘train’, *mees* ‘man’, and *hoov* ‘yard’. Interestingly, only one item, *rong* ‘train’, is among the lemmas for which the corpus-based model has predicted that there is a very high probability of the *peal* construction; for the other two items the corpus-based model has also predicted that a *peal* construction is the preferred one, but the probability estimates are much lower. We suspect that the reason why native speakers prefer the *peal* construction for *rong* 'train' and *mees* 'man' has to do with constructional polysemy.

Since the adessive construction is used to express possession in Estonian, the native speakers have chosen the *peal* construction with the items *rong* 'train' and *mees* 'man' in order to avoid the ambiguity between the possessive and locative reading that is created by using these items in the adessive construction (as in examples (3) and (4)). At the same time, the same construction can also be used in a locative clause, as in example (5). Still, we propose that the default reading of items like *rong* 'train' and *mees* 'man' in the adessive construction is that of possession. For the item *hoov* 'yard', we entertain the explanation that since it is a short, one-syllable word, it prefers the *peal* construction.

(3) ***Rongi-l*** *on* *ratta-d.*
train-SG.ADE be.PRS.3SG wheel-PL.NOM
'The train has wheels.'

(4) ***Mehe-l*** *on* *koer.*
man-SG.ADE be.PRS.3SG dog.SG.NOM
'The man has a dog.'

(5) *Ta* *kohtu-s* *tema-ga* *Tallinna-st* *Tartu-sse*
S/he SG.NOM meet-PST.3SG s/he-SG.COM Tallinn-SG.ELA Tartu-SG.ILL
sõitva-l ***rongi-l.***
going-SG.ADE train-SG.ADE
'He met her on a train going from Tallinn to Tartu'.

There are also items in the first subgroup that show a clear preference for the adessive construction in both tasks: *keldrikorrus* 'basement-level floor', *metsaserv* 'edge of a forest', *krossirada* 'racing course' and *lävi* 'threshold'. The three first items are ranked the lowest for the probability of the *peal* construction also according to the corpus-based model, i.e. the model predicts the adessive construction for these items. Another distinctive characteristic about these three items is that they are compound words and denote static places (i.e. items that cannot be moved or move on their own). The two predictors – morphological complexity (whether the Landmark word is a compound or a simple word) and mobility are the two most significant linguistic predictors that determine the choice between the two constructions according to the corpus-based mixed model. We will come back to this issue in the next section (Section 4.3), where we discuss the role of complexity and other linguistic predictors in determining the choice and preference of one construction over another.

Let us now look at the second group of items for which we have converging evidence across the two sets of experimental data – items for which neither the forced choice data nor the ratings data show a clear preference of one construction over another. This group includes items clustered around zero in Figure 3, for example, *telgilauake* (lit. 'tent-desk'), *lauahunnik* (lit. 'board-pile'), *paber* 'paper', *muld* 'earth', *matt* 'mat', *sõrm* 'finger', *pink* 'bench', *mis* 'what', and *hing* 'soul'. If we look at Figure 1 (the corpus-based probability estimates for the experimental items), we see that none of these items fall in the probability range of 0.4–0.6; this is where we would find items for which both constructions are equally probable according to the corpus model. This means that although the native speakers do not prefer one construction over another for these 9 items, the corpus model does.

Finally, let us look at items for which there is either a clear choice or a clear difference in ratings, but for which the same trend cannot be detected in the other source of data. There are four logical possibilities: (i) items for which there is a clear choice, but no difference in ratings; (ii) items for which there is a clear choice, but the alternative has received a higher rating instead; (iii) items for which there is a clear difference in ratings, but no difference in choices; and (iv) items for which there is a clear difference in ratings, but the proportion of choices is higher for the alternative. Based on the graphs in Figure 4 we conclude that there were no items in our data that fall into categories (ii), (iii) or (iv). There are three items in category (i) – *redel* 'ladder', *praam* 'ferry' and *kivi* 'stone'. The forced choice data indicates a clear preference for the *peal* construction for these three items, but the ratings for these two constructions are virtually the same across the three items. The corpus-based model also predicts the *peal* construction for *kivi* 'stone' and *praam* 'ferry', but for *redel* 'ladder' the corpus assigns a more or less equal probability for both constructions. In general, this result shows that the forced choice data is better aligned with the corpus data than the rating data.

4.3. The role of individual predictors

We saw above that the mixed-model fitted to the corpus data picked up four fixed effects: LM_LENGTH, LM_COMPLEXITY, LM_MOBILITY, and TR_WORDCLASS. The estimated coefficients for the main effects indicate that the adessive construction is preferred when the Landmark is a compound denoting a static place. Conversely, the *peal* construction is preferred when the Landmark is a mobile thing denoted by a short, simple word. Nominal Trajectors predict the adessive, while pronominal and verbal Trajectors predict the *peal* construction. We may inspect how the same four predictors – LM_LENGTH, LM_COMPLEXITY, LM_MOBILITY, TR_WORDCLASS – behave according to the forced choice and ratings data. Figure 5 presents 8 plots: the four plots in the upper row show the proportion of choices for the two constructions across the four predictors; the four plots in the lower row show the residualised mean ratings for the two constructions across the four predictors.

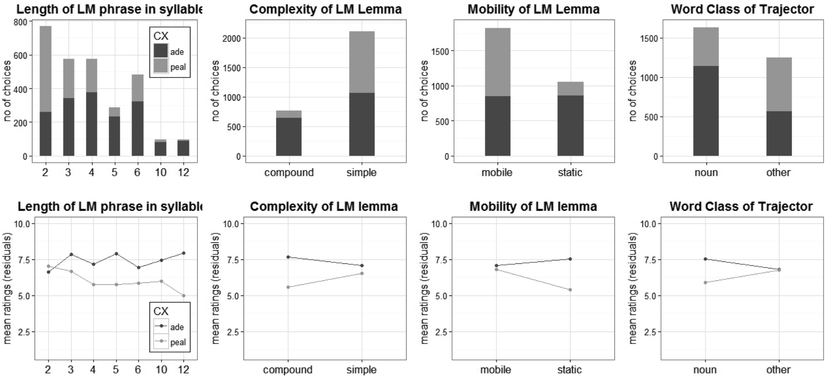


Figure 5. Effect display for the four predictors across the two constructions (dark grey for adessive construction and light grey for *peal* construction) in the forced choice data (upper plots) and ratings data (lower plots).

Several points are worth noting as regards the plots in Figure 5. First of all, as far as the four predictors are concerned, the forced choice data and rating data show similar trends. Moving from left to right we see that the longer the Landmark phrase, the higher the proportion of adessive choices (dark grey in Figure 5) compared to *peal* choices (light grey in Figure 5) and the higher the mean rating for adessive construction compared to *peal* construction. As for complexity of the Landmark lemma, both forced choice data and rating data indicate that while the proportion of choices and the mean rating for the two constructions is virtually equal for simple Landmark lemmas, there is a pronounced difference for compound Landmark lemmas which clearly prefer the adessive construction. Moving on to mobility of the Landmark lemma, we see a pronounced difference between the choices and ratings of the two constructions with static Landmarks, but not with mobile Landmarks. The preferred construction with static Landmarks is the adessive constructions, as with nominal Trajectors. In a nutshell, the plots in Figure 5 mirror the results of the mixed-effects logistic regression model fitted to the corpus data.

The second point that needs to be highlighted in regard with the plots in Figure 5 is that, as far as the rating data is concerned, the *peal* construction seems to be much more sensitive to the changes in the conditions of the four predictors than the adessive construction. What we mean here is that when we take, for example, the lower plots in Figure 5 we see that the adessive construction has received an equally high rating irrespective whether the Landmark is a compound or a simple word, whether it is mobile or static, or whether the Trajector is a noun or belongs to another word class – the position of the dark grey dots remains virtually the same across the different plots. However, the position of the light grey dots varies considerably when we move from one condition to another indicating that the *peal* construction is sensitive to the changes in the conditions of the predictors.

5. Discussion: corpus-based predictions vs preferential choices and ratings

From the analysis of the results, three important conclusions can be drawn: first, the corpus-model predictions correlate well with both the forced choices and ratings (with a slightly better correlation between predictions and choices); second, there is a very strong correlation between the forced choice data and the rating data; and third, the four predictors singled out by the mixed-effects logistic regression model (length, complexity, mobility, word class) behave in a similar way across all three datasets (slightly less for the rating data). However, the most important and crucial conclusion to be drawn from our study is that the default choice for Estonian native speakers is the adessive construction. The reason why we claim that the adessive construction is the default choice for native speakers is based on the results of both the forced choice and the rating experiment. The forced choice experiment shows that, overall, the proportion of the adessive constructions (60%, 1705 out of 2880) was remarkably higher than the proportion of *peal* constructions (40%, 1175 out of 2880). This is mirrored by the rating study, where the overall mean residualised rating was considerably higher for the adessive construction (7.2) compared to the *peal* construction (6.3). It is important to stress that under the null hypothesis, the two constructions should have been chosen in equal proportions and should have received an equally high or low rating, since there were roughly an equal number of sentences where the original construction was the adessive or *peal*. We follow up on the claim that the adessive construction is the default one by discussing it in the context of frequency. In the corpus analysis of the data used for the present study (Klavan 2012), no such general claims about the use of the two constructions could be made since the original model was fed, by design, an equal number of constructions.

A well-known fact about language use concerns human sensitivity to frequency information (Divjak and Caldwell-Harris 2015, Divjak and Gries 2012, Ellis 2002, Gries and Divjak 2012). It is a natural assumption that frequency effects in all their guises (semantic, lexical, morphological, syntactic) may influence the choice between the two constructions. The results of the two experiments conducted with the adessive and *peal* construction demonstrate that the native speakers of Estonian are attuned, at least, to global frequency – the adessive construction is 10 times more frequent than the *peal* construction in the locative function in the corpus of present-day written Estonian (Klavan 2012: 182–183). We therefore assume that it is very likely that the speakers tend to prefer the more frequent construction when they have no preference (cf. also Divjak *et al.* 2016). The overall higher frequency of the adessive construction is the likely culprit why the proportion of adessive choices was higher in the forced choice task and why the mean residualised rating was significantly higher for the adessive construction compared to the *peal* construction. The corpus model, however, had no access to information about the token frequencies of the two constructions, other than their frequencies in the sample, which was equal by design. Divjak *et al.* (2016) show that once their original model was adjusted for frequency, it performed exactly at the same level as the average human participant. The model started to behave more like speakers with overgeneralizations of the most frequent verb. This finding provides support for the assumption that speakers use frequency information and, if possible, this information should be included in the corpus-based models.

The corpus model fitted to the Estonian data does not accommodate frequency information. Operationalization of frequency for the Estonian alternation is not as straightforward as simply counting the number of tokens. One of the factors that complicates the inclusion of frequency is constructional polysemy. In addition to expressing location, the Estonian adessive case also expresses temporal relations, states, possessors in possessive clauses, agents with finite verb forms, instruments, and manner. In fact, it is far more common for the adessive to express temporal and other abstract relations than location (Klavan 2012: 103–108). It is not clear whether native speakers conceive of the different functions of the adessive construction as polysemous or homonymous. This, in turn, has direct consequences of how to count the number of adessive tokens in the corpus sample – should only the locative use of the adessive be taken into account or should all occurrences of the construction be included. The question is clearly empirical in nature, but requires extensive manual annotation and remains, therefore, an undertaking for the future.

In addition to frequency, our study indicates that the four predictors singled out by the mixed-effects logistic regression model fitted to the corpus data show similar trends across the two experiments. Both the forced choice data and the ratings data confirm that longer, complex and static Landmarks prefer the adessive construction, while shorter Landmark phrases prefer the *peal* construction. Nominal Trajector phrases tend to co-occur with the *peal* construction, while other types of Trajector phrases co-occur with the adessive construction. It seems safe to assume that native speakers either consciously or subconsciously make use of at least some of the same distributional aspects as modelled in the corpus data. However, there are two caveats here: first, we should not forget that a large proportion of the variance in the corpus data is explained by the random factors of individual lemmas; second, it may be true that a combination of entirely different set of predictors does an equally good job at predicting the choices and ratings between the two constructions.

As to the first caveat, a number of studies have demonstrated that including lemmas and subjects as random effects in regression models yield a significant improvement in model fit (e.g. Baayen *et al.* 2013, Bresnan and Ford 2010, Janda *et al.* 2012, Theijssen *et al.* 2013). This has raised concerns whether the higher-level abstract features (e.g. syntactic, semantic and discourse-related features) linguists choose to annotate for corpus data are in fact relevant for native speakers when choosing between alternative constructions or lexemes (Divjak 2015, Theijssen *et al.* 2013). The same concern holds for the present data – although both the corpus-based regression model as well as the two experiments indicate that length, complexity, mobility and word class are significant predictors, it is currently only hypothetical. We need psycholinguistic experiments in order to find out whether language speakers actually "think" in terms of these abstract features or categories we have chosen to annotate in our data. Psycholinguistic experiments are also needed to confirm whether the choice between alternative constructions for native speakers is determined by the individual lemmas, as seems to be the case according to the mixed-effects models fitted to the corpus data.

The second caveat concerning the specific predictors picked up by the corpus-based model and the native speakers pertains to the fact that a selection of different predictors may do an equally good job in approximating human behaviour. The study conducted by Divjak *et al.* (2016) demonstrates that a number of models with a random selection of features performs at an equal footing. The authors suggest that "it does not really matter what exactly [language] learners track, as long as they track enough features" (Divjak *et al.* 2016: 29). A similar point is made by Baayen (2011) who shows for a set of models that the overall accuracy is hardly affected by permuting the values of a single predictor. It seems to be the case that individual higher-level abstract features are not that important, which is likely due to the correlational structure of the predictor space (Baayen 2011: 306): any given feature or predictor is predictable from other features or predictors. For the Estonian dataset, for example, COMPLEXITY, LENGTH and MOBILITY are to some extent correlated – compound nouns are longer and denote static places, simple nouns are shorter and denote mobile things.

Rampant collinearity, a common characteristic of language, can pose serious problems for statistical modelling (specifically logistic regression). This has led for the call to prefer statistical modelling techniques that mirror human behaviour when doing quantitative data analysis in linguistics (e.g. Baayen 2011). For example, in case of machine classification, we should prefer modelling techniques that do not pose restrictions on collinear predictors (e.g. naïve discriminative learning). Such modelling techniques are, at least in theory, cognitively more plausible since it is precisely this redundancy that "makes human learning of language data robust" (Baayen 2011: 309) and "explains how individual differences and uniformity across the community can co-exist" (Divjak *et al.* 2016). For the Estonian data, this line of research is under way, but the call for psycholinguistic experimentation as a means of validating the results of a corpus-based model remains, regardless of the modelling technique used.

6. Conclusion

In this paper, we compared the forced choice data and the rating data against a mixed-effects logistic regression model fitted to the corpus data of the alternation between the adessive case and the adposition *peal* 'on' in present-day written Estonian. We show that there is a strong positive correlation between the corpus-based probability estimates and the experimental data – as the probability of the *peal* construction rises, so does the proportion of choices and the mean residualised rating for the *peal* construction. Furthermore, we demonstrate that the two sources of experimental data, by and large, converge – choices and ratings correlate very strongly. At the same time, our study also shows that there are instances where the two sources of data diverge – for certain experimental items there was a clear choice, but the ratings for the two constructions were virtually the same. In addition, the rating data showed that while the *peal* construction is sensitive to the changes in the conditions across the linguistic predictors, the adessive construction retains a high rating across the conditions of the predictors. We therefore conclude that both sources of data are necessary as they provide important complementary information as to the nature of the alternating constructions.

Generalising over the results of the corpus-based model and the experimental data, we have found confirmation that compound nouns denoting static places predict the adessive construction, while shorter words denoting small, manipulable or movable objects predict the *peal* construction (see also Bartens 1978, Klavan 2012, Klavan *et al.* 2015, Ojutkangas 2008). Although our study shows that the forced choice data and the rating data mirror the results of the corpus-based regression model, we cannot be 100% sure that speakers actively tap into the predictors singled out by the corpus model when choosing or rating the two constructions. We may need to test individual predictors to verify whether in fact these predictors play the crucial role as the corpus and experimental data seem to exhibit. Another conclusion that needs further verification is the fact that a large proportion of the variation seems to be explained by individual lemmas.

One of the crucial points that our study makes is that experimental data highlights the importance of constructional frequency. The default construction for native speakers in this specific alternation is the adessive constructions. The choice for the adessive construction was proportionally higher than for the *peal* construction in the forced choice task and the mean residualised rating for the adessive construction was much higher than the mean residualised rating for the *peal* construction. Although the adessive construction is much more frequent in present-day written Estonian (and this is reflected in the experimental data, we believe), constructional frequency has not been taken into account in the corpus-based model used in the present study. Future corpus-based work needs to account for frequency in order to provide a cognitively plausible model of the Estonian data.

Acknowledgements

We are grateful to the anonymous referees for the journal and to Liina Lindström in particular for their meticulous comments on this paper. The views expressed and the responsibility for any errors remains our own. The work of the first author was supported by a grant from the Estonian Research Council (PUT1358 "The Making and Breaking of Models: Experimentally Validating Classification Models in Linguistics").

Addresses:
Jane Klavan
Department of English Studies
University of Tartu
Lossi 3
51003 Tartu, Estonia
E-mail: Jane.Klavan@ut.ee

Ann Veismann
Institute of Estonian and General Linguistics
University of Tartu
Jakobi 2
51014 Tartu, Estonia
E-mail: Ann.Veismann@ut.ee

References

Arppe, Antti and Dana Abdulrahim (2013) “Converging linguistic evidence on two flavors of production: the synonymy of Arabic COME verbs”. Paper presented at Second Workshop on Arabic Corpus Linguistics, University of Lancaster, 22–26 July.

Arppe, Antti, Patrick Bolger, and Dagmara Dowbor (2012) “The more evidential diversity, the merrier – contrasting linguistic data on frequency, selection, acceptability and processing”. Paper presented at New Ways of Analyzing Syntactic Variation (NWASV), Radboud University, Nijmegen, the Netherlands, 2012.

Arppe, Antti and Juhani Järvikivi (2007) “Every method counts: combining corpus-based and experimental evidence in the study of synonymy”. *Corpus Linguistics and Linguistic Theory* 3, 2, 131–159.

Baayen, R. Harald, Anna Endresen, Laura A. Janda, Anastasia Makarova, and Tore Nesset (2013) “Making choices in Russian: pros and cons of statistical methods for rival forms”. *Russian Linguistics* 37, 253–291.

Baayen, R. Harald (2011) “Corpus linguistics and naive discriminative learning”. *Brazilian Journal of Applied Linguistics* 11, 295–328.

Baayen, R. Harald, Douglas J. Davidson, and Douglas M. Bates (2008) “Mixed-effects modeling with crossed random effects for subjects and items”. *Journal of Memory and Language* 59, 390–412.

Bartens, Raija (1978) *Synteettiset ja analyyttiset rakenteet lapin paikanilmauksissa* (Suomalais-ugrilaisen Seuran toimituksia, 166.) Helsinki: Suomalais-Ugrilainen Seura.

Bermel, Neil and Luděk Knittl (2012) “Corpus frequency and acceptability judgments: a study of morphosyntactic variants in Czech”. *Corpus Linguistics and Linguistic Theory* 8–2, 241–275.

Bresnan, Joan (2007) “Is syntactic knowledge probabilistic? Experiments with the English dative alternation”. In Sam Featherston and Wolfgang Sternefeld, eds. *Roots: linguistics in search of its evidential base*, 77–96. Berlin: Mouton de Gruyter.

Bresnan, Joan and Marilyn Ford (2010) “Predicting syntax: processing dative constructions in American and Australian varieties of English”. *Language* 86, 1, 186–213.

Burnham, Kenneth P. and David R. Anderson (2002) *Model selection and multimodel inference: a practical information-theoretic approach*. 2nd ed. New York: Springer.

Comrie, Bernard (1986) “Markedness, grammar, people, and the world”. In Fred R. Eckman, Edith A. Moravcsik, and Jessica R. Wirth, eds. *Markedness*, 85–106. New York: Plenum.

Crawley, Michael J. (2007) *The R book*. Chichester: John Wiley & Sons.

Divjak, Dagmar (2016) “The role of lexical frequency in the acceptability of syntactic variation. Evidence from that-clauses in Polish”. *Cognitive Science*, 1–29.

Divjak, Dagmar (2015) “Exploring the grammar of perception. A case study using data from Russian”. *Functions of Language* 22, 1, 44–68.

Divjak, Dagmar, Antti Arppe, and Harald R. Baayen (2016) “Does language-as-used fit a self-paced reading paradigm? (The answer may well depend on the statistical model you use)”. In Anja Gattnar, Tanja Anstatt, and Christina Clasmeier, eds. *Slavic languages in the black box*, 52–82. Tübingen: Narr-Verlag.

Divjak, Dagmar, Antti Arppe, and Ewa Dąbrowska (2016) “Machine meets man: evaluating the psychological reality of corpus-based probabilistic models”. *Cognitive Linguistics* 27, 1, 1–33.

Divjak, Dagmar and Catherine L. Caldwell-Harris (2015) “Frequency and entrenchment”. In Ewa Dąbrowska and Dagmar Divjak, eds. *Handbook of cognitive linguistics,* 53–75. Berlin: Mouton de Gruyter.

Divjak, Dagmar and Stefan Th. Gries (2012) *Frequency effects in language representation*. Volume 2. (Trends in linguistics. Studies and monographs, 244.2). Berlin: Mouton de Gruyter.

Divjak, Dagmar and Stefan Th. Gries (2008) “Clusters in the mind? Converging evidence from near-synonymy in Russian”. *The Mental Lexicon* 3, 2, 188–213.

Ellis, Nick C. (2002) “Frequency effects in language processing”. *Studies in Second Language Acquisition* 24, 2, 143–188.

Erelt, Mati, Reet Kasik, Helle Metslang, Henno Rajandi, Kristiina Ross, Henn Saari, Kaja Tael, and Silvi Vare (1995) *Eesti keele grammatika I. Morfoloogia*. Tallinn: Eesti Teaduste Akadeemia Eesti Keele Instituut.

Erelt, Mati, Tiiu Erelt, and Kristiina Ross (2007) *Eesti keele käsiraamat*. Tallinn: Eesti Keele Sihtasutus.

Ford, Marilyn and Joan Bresnan (2013) “Using convergent evidence from psycholinguistics and usage”. In Manfred Krug and Julia Schlüter, eds. *Research methods in language variation and change*, 295–312. Cambridge: Cambridge University Press.

Gilquin, Gaëtanelle and Stefan Th. Gries (2009) “Corpora and experimental methods: a state-of-the-art review.” *Corpus Linguistics and Linguistic Theory* 5, 1, 1–26.

Gries, Stefan Th. and Dagmar S. Divjak (2012) *Frequency effects in language learning and processing*. Vol. 1. (Trends in linguistics. Studies and monographs, 244.1). Berlin: Mouton de Gruyter.

Hagège, Claude (2010) *Adpositions*. Oxford: Oxford University Press.

Harrell, Frank E. (2001) *Regression modeling strategies: with applications to linear models, logistic regression and survival analysis.* New York: Springer.

Hosmer Jr, David W., Stanley Lemeshow, and Rodney X. Sturdivant (2013) *Applied logistic regression*. Hoboken, NJ: John Wiley & Sons.

Janda, Laura A. (2013) “Quantitative methods in *cognitive linguistics*: an introduction”. In Laura A. Janda, ed. *Cognitive linguistics: the quantitative turn. The essential reader*, 1–9. Berlin: Mouton de Gruyter.

Janda, Laura, Svetlana Sokolova, and Olga Lyashevskaya (2012) “The locative alternation and the Russian ‘empty’ prefixes: a case study of the verb gruzit’ ‘load’”. In Dagmar Divjak and Stefan Th Gries, eds. *Frequency effects in language representation,* 51–86. (Trends in linguistics. Studies and monographs, 244.2). Berlin: Mouton de Gruyter.

Klavan, Jane (2012) *Evidence in linguistics: corpus-linguistic and experimental methods for studying grammatical synonymy*. (Dissertationes linguisticae universitatis Tartuensis, 15). Tartu: University of Tartu Press.

Klavan, Jane and Dagmar Divjak (2016) “The cognitive plausibility of statistical classification models: comparing textual and behavioral evidence”. *Folia Linguistica* 50, 2, 355–384.

Klavan, Jane, Maarja-Liisa Pilvik, and Kristel Uiboaed (2015) “The use of multivariate statistical classification models for predicting constructional choice in spoken, non-standard varieties of Estonian”. *SKY Journal of Linguistics* 28, 187–224.

Langacker, Ronald W. (2008) *Cognitive grammar. a basic introduction.* Oxford: Oxford University Press.

Lestrade, Sander (2010) *The space of case.* Unpublished doctoral dissertation. Radboud University Nijmegen.

McCulloch, Charles E. and John M. Neuhaus (2011) “Misspecifying the shape of a random effects distribution: why getting it wrong may not matter”. *Statistical Science* 26, 3, 388–402.

Neuhaus, John M., Charles E. McCulloch, and Ross Boylan (2013) “Estimation of covariate effects in generalized linear mixed models with a misspecified distribution of random intercepts and slopes”. *Statistics in Medicine* 32, 14, 2419–2429.

Ojutkangas, Krista (2008) “Mihin suomessa tarvitaan sisä-grammeja?”. *Virittäjä* 3, 382–400.

Pinheiro, José C. and Douglas M. Bates (2000) *Mixed-effects models in S and S-PLUS.* New York: Springer.

Roland, Douglas, Jeffrey L. Elman, and Victor S. Ferreira (2006) “Why is *that*? Structural prediction and ambiguity resolution in a very large corpus of English sentences”. *Cognition* 98, 245–272.

Theijssen, Daphne, Louis ten Bosch, Lou Boves, Bert Cranen, and Hans van Halteren (2013) “Choosing alternatives: using Bayesian networks and memory-based learning to study the dative alternation”. *Corpus Linguistics and Linguistic Theory*, 227–262.

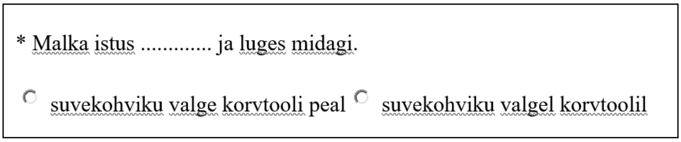
Wasow, Thomas and Jennifer Arnold (2003) “Post-verbal constituent ordering in English”. *Topics in English Linguistics* 43, 119–154.

Kokkuvõte. Jane Klavan ja Ann Veismann: Kas keelekasutajate valikud ja hinnangud peegeldavad korpuspõhiseid tõenäosushinnanguid? Eesti keele adessiivi ja kaassõna *peal* kasutus tänapäeva kirjakeeles. Tänapäevases kasutuspõhises keeleteaduses rõhutatakse vajadust kombineerida korpusandmetele toetuvat analüüsi katseliste uuringutega. Mitmed uurimused on võrrelnud korpusel põhineva statistilise mudeli headust emakeelsete kõnelejate käitumisega keelelistes katsetes. Käesolev artikkel jätkab seda uurimisliini, pannes võrdlusesse korpusandmetega kaks keelelist katset. Artiklis hinnatakse korpuspõhise segamudeli headust, võrreldes seda eesti keelt emakeelena kõnelejate käitumisega sunnitud valiku katses ja hinnangukatses. Uuritavaks nähtuseks on eesti keele adessiivi ja kaassõna *peal* rööpne kasutus kohasuhete väljendamisel tänapäeva kirjakeeles.

Võtmesõnad: konstruktsioonilised alternatsioonid, korpuslingvistika, sunnitud valiku katse, hinnangukatse, statistiline mudeldamine, eesti keel

Appendix 1: Sample items for the two experiments

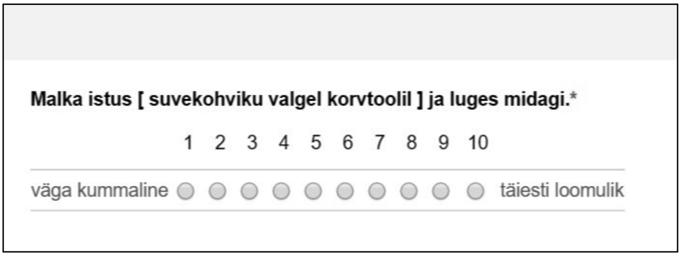
A. Sample item for the forced choice task



* Malka istus .... ja luges midagi.

○ suvekohviku valge korvtooli peal ○ suvekohviku valgel korvtoolil

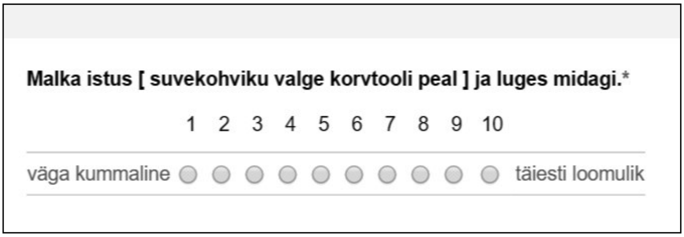
B. Sample item for the rating task (adessive construction)



Malka istus [suvekohviku valgel korvtoolil] ja luges midagi.*

	1	2	3	4	5	6	7	8	9	10	
väga kummaline	○	○	○	○	○	○	○	○	○	○	täiesti loomulik

C. Sample item for the rating task (*peal* construction)



Malka istus [suvekohviku valge korvtooli peal] ja luges midagi.*

	1	2	3	4	5	6	7	8	9	10	
väga kummaline	○	○	○	○	○	○	○	○	○	○	täiesti loomulik

Appendix 2: The full list of experimental items together with the sentence context, the lemma, its English translation equivalent, and corpus probability (in ascending order of corpus-based probabilities from very low probability (0) to very high probability (1) of the *peal* construction)

Item	Original sentence	Lemma	Corpus probab.
1	Tänavanurgal silmasid nad, mida neil vaja: **keldrikorrusel** oli leivapood.	*keldrikorrus* 'basement floor'	0.07
2	Ainult kiidusõnu väärib "Osooni" operaator Raul Priks, kelle kaamera oli kinni püüdnud nii **metsaserval** teed kalpsavad kitsed kui ka lume alt toidupoolist toidu otsiva metskitse.	*metsaserv* 'forest edge'	0.09
3	Malka istus **suvekohviku valgel korvtoolil** ja luges midagi.	*korvtool* 'wicker chair'	0.09
4	Kristers Sergis teostab **krossirajal** isa unistust.	*krossirada* 'off-road circuit'	0.09
5	Nad pole enam asjad iseeneses, pole lihtsalt see või teine kivike, see või teine maasikapeenar, see või teine kriimustus **autouksel**, see või teine nihe tapeedimustris.	*autouks* 'car door'	0.09
6	Inimkultuuri osad installatsioonides on näiteks saabas Läänemerest, kohustusliku õnne aeg jõulud, lapsepõlve lahutamatu osa täheklotsid, vampiiri hambad ookeanirannikult ja igapäevane ilm **samal leheküljel** päeva olulisemate juubilaridega.	*lehekülg* 'page'	0.12
7	Päeval oli seal nii kuum, et steariinküünal sulas **telgilauakesel** loiguks.	*telgilauake* 'tent desk'	0.14
8	Paari päeva pärast märkas Leopold äkki ateljee aknast, et **tänaval** jalutab Liis, kõnnib aeg-ajalt maja poole vaadates mööda, siis jälle tagasi.	*tänav* 'street'	0.14
9	Miisu on **katusel**.	*katus* 'roof'	0.16

Item	Original sentence	Lemma	Corpus probab.
10	Temaga oldi lepitud kui paratamatusega, sest olnuks tõepoolest mõttetu nõuda, et ta veel **vanaduse lävel** uue ameti selgeks õpiks.	*lävi* 'threshold'	0.25
11	Pärast istusime maja ees **lauahunnikul**, jõime pikkadest kandilistest pudelitest õlut, sõime kaasavõetud võileibu, vahtisime heldinult maja veel tühje aknaauke ja ma otsustasin, et hakkan aegsasti uusi kardinaid heegeldama.	*lauahunnik* 'pile of boards'	0.26
12	Ehkki on möödas juba seitsesada aastat päevast, mil Aleksander Suur siinsamas Issose lahe ääres ja **Pinarose jõel** purustas Pärsia kuninga Dareiose sõjaväe, muutub mälestus sellest kohe elavaks, kui oled astunud Arese templisse.	*jõgi* 'river'	0.27
13	Aga siis seletas Leonid, kes ühel õhtul saunast tulles tavalisest jutukam oli, et naiste lahkhelid olevat tekkinud **puhtmoraalsel pinnal**.	*pind* 'level'	0.36
14	Viimased viis aastat istub ta **ETTK tegevdirektori toolil**.	*tool* 'chair'	0.42
15	Minna oleks sauna tagant heinamaa servast hea meelega paari sao jagu oma napile loomatoidule lisa niitnud ja küllap ta selle loo kuidagi **redelite peal** kuivaks ka oleks saanud, kuid ei julgenud.	*redel* 'ladder'	0.55
16	"Õhtul hiljem vaatasin **diivani peal** telekat, kui äkki käis uks -- piraki ! -- lahti ja sisse tulid kaks meest, mustad maskid üle pea, silmaaugud sees, nagu K-Komando," kirjeldab Ivar.	*diivan* 'couch'	0.57
17	**Ühe pildi peal** istus mitu naist ümber väikese laua, kohvitassid ees, kübarad peas, rebasenahad kaelas ja kõrge kontsaga kingad jalas.	*pilt* 'picture'	0.67
18	Eks sa katsu selliseid asju **paberi peal** tõestada – ei õnnestu.	*paber* 'paper'	0.67

Item	Original sentence	Lemma	Corpus probab.
19	Ma ei näinud muud kui musta mulda põõsa all ja kirjusid kanasulgi **mulla peal**, päike paistis silma, kui pead tõstsin, pidin vahetevahel silmi hõõruma.	*muld* 'soil'	0.71
20	**Selle hoovi peal** pole võimalik riideid puhtana hoida.	*hoov* 'yard'	0.74
21	Kuulid olid teda tabanud kõhtu, reide ja jalga, ta lamas ristseliti **teise**, **surnud mehe peal**.	*mees* 'man'	0.74
22	Kanti raamat oli tal haiglas ja Irus kogu aeg **voodi peal**.	*voodi* 'bed'	0.79
23	Õhtuks oli vana pruun riidekapp **praami peal** ja sõitis üle mere.	*praam* 'ferry'	0.80
24	"Jalutame sinna eemale, seal **kivide peal** on mõnus istuda, hästi ilus," tegi ta ettepaneku.	*kivi* 'stone'	0.81
25	Ma lähen uitan **rongi peal**, nüüd ma tean siia tagasi tulla.	*rong* 'train'	0.87
26	Naised tõstavad olümpial kangi ja annavad teineteisele **mati peal** kolki, aga miks mehed ei siruta oma karvaseid jalgu medalite eest vee seest välja!	*matt* 'mat'	0.87
27	Masinad pidid kogu töö ära tegema ja inimene pidi ülearuseks muutuma, kui ta enam vikatiga ei vehi ja **sõrmede peal** ei arvuta.	*sõrm* 'finger'	0.87
28	Ta istus siin majaesisel väikesel platsil **pingi peal** koos ühe omavanuse poisiga ja vestles sellega elavalt.	*pink* 'bench'	0.88
29	Tuvi selle peale ei mõtle, kelle pea see on, **mille peal** on mõnus istuda.	*mis* 'what'	0.88
30	Tal on kindlasti midagi **hinge peal**, mida ta ainult meile usaldab.	*hing* 'soul'	0.89