Jane Klavan, Maarja-Liisa Pilvik and Kristel Uiboaed[1]

The Use of Multivariate Statistical Classification Models for Predicting Constructional Choice in Spoken, Non-Standard Varieties of Estonian[2]

Abstract

Speakers' choice between alternative forms is driven by a variety of factors. Multivariate statistical classification models enable us to discover which factors are significant and to what extent. In this paper we take a close look at two near-synonymous constructions – the synthetic adessive construction (e.g. *laual* 'on the table') and the analytic adpositional *peal* 'on' construction (*laua peal* 'on the table') – expressing locative function in Estonian, and identify a number of semantic and morpho-syntactic factors that influence the choice between the two constructions. In the first systematic study on the subject (Klavan 2012) a logistic regression model with four morphosyntactic and two semantic explanatory predictors was fit to Estonian written language data, yielding a classification accuracy of 70%. In our study, we use dialectal data from the Corpus of Estonian Dialects (CED 2015) to explore how the minimal adequate model for written data performs on non-standard, spoken spontaneous language. In addition, we include the geographical dimension and the Landmark lemma as random effects and demonstrate how these factors significantly improve the fit of the model. Furthermore, we show how complementing the results of the mixed-effects logistic regression model with the results obtained with the 'tree & forest' models helps to explain the variation in more detail and highlight significant interactions in the data.

1. Introduction

A common presumption in usage-based linguistics is that speakers' linguistic knowledge is probabilistic in nature. It has been shown that speakers have a richer knowledge of linguistic constructions than the

[1] The authors are listed in alphabetical order.

[2] This study was supported by Estonian Research Council (PUT 90 Estonian Dialect Syntax).

knowledge captured by categorical judgements leads us to believe (e.g. Bresnan 2007; Bresnan, Cueni, Nikitina & Baayen 2007; Bresnan & Ford 2010; Tagliamonte & Baayen 2012; Divjak & Arppe 2013; Szmrecsanyi 2013). In addition to the probabilistic nature of linguistic data, language use is also driven by multitude of factors. Speakers' choice between alternative forms is often influenced by semantic, syntactic, morphological, phonological, discourse-related, lectal, and other factors. The practical and methodological question is how can we capture this knowledge quantitatively. At the moment, multivariate statistical classification modeling seems to be the best tool available. The present paper continues this line of research and discusses the results of a multivariate corpus analysis of two near-synonymous constructions in Estonian. We take a usage-based and variationist perspective and focus on non-standardized, spoken spontaneous language. We look at the parallel use of the synthetic adessive case construction and the analytic adposition *peal* 'on' construction in Estonian dialects.

The paper has two overall aims. The first aim is to compare how the model fitted to the dialect data performs in comparison to the model fitted to the written language data. To this end a multivariate corpus analysis was carried out with 2,032 occurrences of the adessive case and the adposition *peal* 'on' in the Corpus of Estonian Dialects (CED 2015[3]). The dialect data were analysed using mixed-effects logistic regression (Pinheiro & Bates 2002; Baayen, Davidson & Bates 2008). The minimal adequate model fitted to the written language includes four morphosyntactic and two semantic explanatory predictors and has a classification accuracy of 70% (Klavan 2012). We are interested in testing whether the same morphosyntactic and semantic predictors are also significant for predicting the choice in non-standard spoken language. We are furthermore interested to see whether the fit of the model can be significantly improved by including the geographical dimension in the model. It has been suggested that the use of analytic locative constructions (in our case the adposition *peal* constructions) are more characteristic of Southern and Western Estonia, while the use of synthetic locative constructions (in our case the adessive case constructions) are more frequent in Northern Estonia (Palmeos 1985: 15).

The second objective of the study is to compare the performance of the logistic regression model with the results obtained using a different

[3] Data were collected in October 2014.

method for prediction – the ‘tree & forest’ method (Breiman 2001; Strobl, Malley & Tutz 2009) – on dialect data in a similar fashion to Tagliamonte and Baayen (2012) and Baayen, Endresen, Janda, Makarova and Nesset (2013). In their comparative studies, Baayen and colleagues conclude that alternative modelling techniques perform best at different aspects of statistical modelling, such as accounting for interactions, classification accuracy etc., and that using the ‘tree & forest’ method to complement logistic models is highly beneficial (Tagliamonte & Baayen 2012: 23; Baayen et al. 2013: 285–287). In this paper we advocate the same methodological pluralism and show on a dataset that is typologically different from the English word order alternations how a combination of logistic regression and the ‘tree & forest’ method enables us to account for the variation between the adessive and *peal* construction in Estonian dialect data.

The rest of the paper is structured as follows: In Section 1.1. we give a short overview of the adessive case construction and the adpositional *peal* construction, followed by a brief general discussion on the rivalry between analytic and synthetic constructions in Estonian (Section 1.2.). The dataset we are working with is introduced in Section 2. The mixed-effects logistic regression model is presented in Section 3, followed by the ‘tree & forest’ modelling of the data in Section 4. The paper ends with a general discussion of the results (Section 5) and conclusions (Section 6).

1.1 Estonian adessive case and adposition *peal* ‘on’

Examples 1 and 2 illustrate how Estonian has at least two different ways to express a situation where an object is located on top of another object (the support-relation). The first option is to use the adessive case, as *laual* in example (1) (referred to as the synthetic locative construction). The second option is to use the genitive case together with the adposition *peal* ‘on’, as *laua peal* in example (2) (the analytic locative construction).

(1)	*Raamat*	*on*	***laual***.
	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**.’

In addition to expressing location, the Estonian adessive case also expresses temporal relations (e.g. *neljapäeva**l*** 'on Thursday'), states (e.g. *naeru**l** näoga* 'with a smiling face'), possessors in possessive clauses (e.g. *Mari**l** on kaks last* 'Mari has two kids'), agents with finite verb forms (e.g. *See asi ununes mu**l** kiiresti* 'I quickly forgot about that thing'), instruments (e.g. *klaveri**l*** 'on a piano'), and manner (e.g. *kikkis kõrvu**l*** 'with ears pricked up') (Erelt, Erelt & Ross 2007: 250). It is far more frequent for the adessive to express temporal and other abstract relations than location (Klavan 2012: 103–108). Similarly to locative cases, Estonian adpositions are also polysemous. The Dictionary of Written Estonian (Langemets, Tiits, Valdre, Veskis, Viks & Voll 2009: 130–131) lists as many as 11 meanings for the adposition *peal*; relevant for the present study is the locative function given in example (2) above. 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, Kasik, Metslang, Rajandi, Ross, Saari, Tael & Vare 1995: 33–34; Erelt et al. 2007: 191). This is in line with the general claims made in literature concerning the differences between adpositions and case affixes (Comrie 1986; Hagège 2010; Lestrade 2010).

1.2 Analyticity and syntheticity in Estonian

The alternation between synthetic and analytic locative constructions is a typologically intriguing language phenomenon. Although analytic and synthetic rivalry on a more broader scale has attracted a fair amount of interest in usage-based linguistics (cf. Szmrecsanyi 2009, 2012), it has not been extensively studied in Estonian linguistics (save for a few small scale studies, e.g. Serebrennikov 1959; Metslang 1997, 2009). In addition to the alternation between locative cases and the corresponding adpositions, there are other constructions where both the analytic as well as the synthetic alternative are used side by side. Standard Estonian is claimed to have become more analytic due to both its internal development as well as the influence of Russian, English and German (Erelt & Metslang 1998: 659; Metslang 2009: 49). We have even less information about the nature and extent of the analytic and synthetic rivalry in spoken language and nonstandard varieties. Analyticity is claimed to be the influence of the Germanic or Baltic languages (Serebennikov 1959; Erelt & Metslang 1998), which leads us to expect that analytic constructions are more widespread, for example, in areas which have had direct contacts with

Swedish (e.g. Western and Northern-Estonia), whereas the Eastern and Southern part of the country may use synthetic constructions more in the same semantic and morphosyntactic environment. These kinds of tendencies are found in the variation of certain verbal constructions (Uiboaed, Hasselblatt, Lindström, Muischnek & Nerbonne 2013; Uiboaed 2013). However, we lack systematic empirical studies on the analytic and synthetic alternation in the standard and non-standard varieties of the Estonian language. Our paper aims to fill this gap by taking a multivariate quantitative look at one specific alternation between synthetic and analytic constructions in non-standard, spoken Estonian. It continues the work of Klavan (2012) who applied multivariate statistical modelling to study the use of the adessive and the *peal* construction in Estonian written language.

In other Finno-Ugric languages, Bartens (1978) and Ojutkangas (2008) have looked at the alternation between the interior locative cases and the corresponding adpositions in the Saami and Finnish languages respectively. The central claim of Bartens (1978) is that the analytic adpositional construction places more stress on the location than the synthetic case construction. In addition, Bartens (1978) specifies that in the Saami languages the adpositional constructions are used together with smaller, manipulable things as Landmarks[4] as well as with vehicles. Ojutkangas (2008: 386–389) reports the results of a corpus study based on Finnish dialects and her central claim is that the interior locative cases express conventional spatial relations between Trajector and Landmark, while the corresponding adpositional constructions are used when this relation is somewhat unconventional or unexpected for the speaker.

2. Description of the dataset

2.1 Data extraction

A database of 2,032 instances of the adessive and *peal* constructions was collected from the morphologically annotated part of Corpus of Estonian Dialects (CED 2015). At the time of data collection in October 2014 the

[4] We adopt Langacker's (2008) terminology to refer to the two most fundamental notions in relational expressions. The most prominent participant is called Trajector and the second participant Landmark (Langacker 2008: 70). 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 Trajector is specified. Trajector may be static or dynamic, a person or an object, or even a whole event.

corpus consisted of 834,311 morphologically annotated tokens in total from ten traditional dialect areas. CED consists of transcriptions of spoken spontaneous language. Recordings (most of them recorded during 1960s–1970s) are traditional dialect interviews where informants talk about past events, customs, work, and their everyday lives. Texts contain long monologous passages produced by the informants. The selection of informants follows the traditional criteria: not highly educated, non-mobile, elderly people. Our data comes with the traditional problems of variationist and sociolinguistic data. The data are unevenly distributed across dialect areas and individuals, informants are mostly women, we do not have completely comparable linguistic contexts across all the regions, etc. (cf. Tagliamonte & Baayen 2012: 142–143). On the upside, we have a fairly balanced dataset diachronically as most of the interviews remain within the relatively short timespan of 20 years.

The written language dataset of Klavan (2012) consisted of 450 randomly selected observations per construction; for the dialect data we decided to extract all of the occurrences of the two constructions in the corpus in order to achieve maximal geographical representativeness. Research data are extracted only from the informants' texts. The data are collected automatically with R script (version 3.1.2, R development core team 2014). We have extracted all the instances of the adessive case and the adposition *peal* with the symmetrical context span of 10 words. All the context units are collected within one informant's utterance. This extraction procedure resulted in a total of 14,710 observations of the adessive case construction and 1,586 of the adpositional *peal* construction. The database of 16,296 observations was manually coded for the type of function expressed by the constructions. For the present study, we only look at instances where the two constructions expressed a locative function since this is the function in which the alternation between the two constructions is evident the most. After the initial coding stage we were left with 722 observations of the adessive case and 1,310 instances of the *peal construction*, a total of 2,032 observations. Unlike the balanced structure of the written language data sample, the dialect data has a preponderance of *peal* constructions (64%).

2.2 Operationalisation and selection of predictors

The written language dataset includes 900 observations sampled from the Corpus of Written Estonian which were coded for 20 semantic and morpho-syntactic variables with 66 distinct variable categories (Klavan 2012: 70–92). For the present study we focus on only those 6 predictors that were retained in the final, minimal model fitted to the written language data: mobility of Landmark, verb group, length of the Landmark phrase in syllables (logarithmically transformed), morphological complexity of Landmark, word class of Trajector, relative position between Trajector and Landmark. The predictor 'mobility of Landmark' is closely related to the predictor 'type of Landmark'; both predictors were annotated for the present dataset. In addition, we annotated the dialect data for the variables dialect, informant, and lemma. Following is a description of how the predictors were operationalised and the specific predictions made.

Type and mobility of Landmark. Landmarks were coded as either 'mobile' or 'static'. Mobile Landmarks are those that do not have a fixed position in the environment, either because they move by themselves (e.g. humans, animals) or can be moved by an external agent (e.g. a table). Static Landmarks have a fixed position in the environment (e.g. street, market). This predictor is closely related to another predictor included in the original annotation schema – type of Landmark. The latter refers to the general distinction between easily manipulable objects or 'things' (e.g. a sleigh) and large static objects or 'places' (e.g. shore). Bartens (1978) demonstrated that in the Saami languages the synthetic constructions are used when Landmark is a place and analytic constructions are more frequent with things as Landmarks. The same has been demonstrated by Ojutkangas (2008), who studied the use of the interior locative cases and the corresponding adpositions in Finnish. There seems to be ample reason to suspect that the type and mobility of Landmark play a role in the alternation between locative cases and adpositions. More specifically – larger, static locations such as places should lend themselves more easily for abstraction and hence are more likely to be used with the adessive (cf. e.g. Bartens 1978), while as small manipulable (or movable) objects or things prefer the adposition *peal* since adpositions are more concrete and specific than cases and they convey the meaning of spatial location of an object more clearly (Bartens 1978; Palmeos 1985: 18; Comrie 1986; Ojutkangas 2008; Hagège 2010: 37–38; Lestrade 2010).

Verb lemma and verb group. The predominant locative construction in Estonian tends to use a simple copula for expressing location (e.g. *olema* 'to be'), but there are other verbs that can be used together with either the adessive or the adposition *peal* 'on' (e.g. *istuma* 'to sit'). In the initial coding stage, each observation is coded for the verb lemma used with either the adessive or the *peal construction*. Different verbal compounds (particle verbs, modal verb constructions etc.) are coded only for the main verb. In the next stage, the verbs are subcategorised into different groups based largely on Levin (1993) and include the following: 'action' (e.g. *avama* 'open'), 'existence' (e.g. *olema* 'be'), 'motion' (e.g. *jooksma* 'run'), and 'posture' (e.g. *istuma* 'sit'). This predictor also has the level of 'no verb' – this is used for elliptical usages where no overt verb lemma is expressed. The latter is highly characteristic of spoken language.

Landmark length. Length is reported as one of the most crucial variables in numerous studies on various syntactic alternation phenomena (e.g. Hawkins 1994, 2004; Wasow 1997, 2002; Arnold, Wasow, Losongco & Ginstrom 2000; Bresnan et al. 2007; Anttila, Adams & Speriosu 2010). In many cases length is discussed under the headings of weight or complexity. Following Mondorf (2003: 253) and Rohdenburg (2003: 205) we can predict that the analytic adpositional construction will be used in cognitively more demanding environments. Rohdenburg's complexity principle (Mondorf 2003: 294; Rohdenburg 2003) states that "in the case of more or less explicit constructional alternatives, the more explicit option(s) will tend to be preferred in cognitively more complex environments" (Rohdenburg 2003: 205).

There are different ways how weight can be measured. Most studies count the number of words (e.g. Rosenbach 2005; Hinrichs & Szmrecsanyi 2007), but others define it in terms of number of syllables (e.g. McDonald, Bock & Kelly 1993), character counts (Wolk, Bresnan, Rosenbach & Szmrecsanyi 2013: 394–395) or in terms of phonological complexity (e.g. the number of lexical stresses as in Anttila et al. 2010). In the annotation schema of Klavan (2012), the length of the Landmark phrase was measured in both syllables and words. In the present study length is only measured in syllables, since this measure produced a model with a better fit than length measured in words. An important methodological question concerns whether to include the monosyllabic adposition *peal* in the count as a separate syllable or not. Hinrichs and Szmrecsanyi (2007: 453) and Rosenbach (2005: 623) indicate in their studies on the English genitive alternation that the definite or indefinite articles determining the possessed phrase of an *of*-genitive were not included because it provides a natural imbalance and skews the results. For similar reasons, it was decided not to include the adposition *peal* in the counts. Hence, Landmark phrases such as *laual* [*laud*+ADE] and *laua peal* [*laud*+GEN *peal*] were both considered as one word long and disyllabic. A logarithmic transformation is applied to the counts of syllables in order to compress extreme values and reduce skewness.

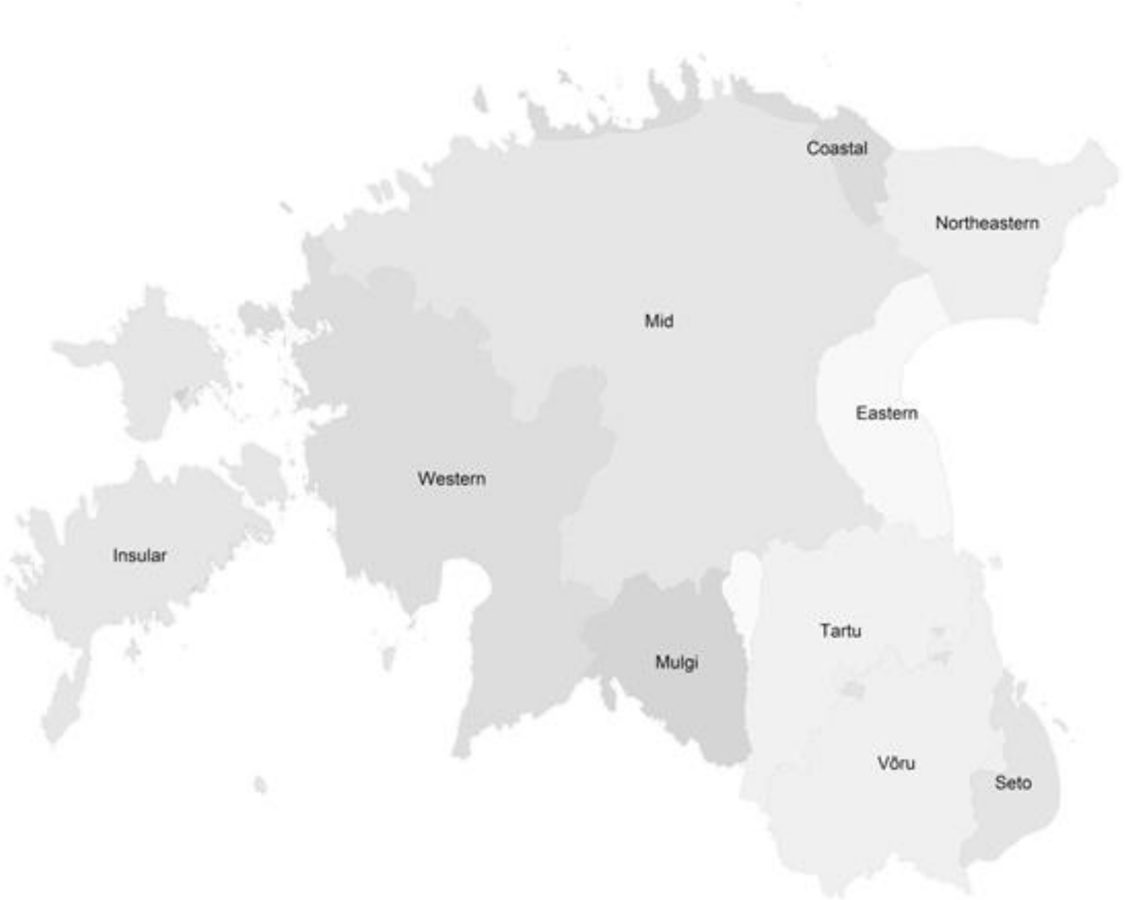
Landmark complexity. It is certainly true that complexity and length are correlated and that longer constituents tend to be more complex and vice versa, but it is still a matter of controversy whether they are two distinct factors or not (Rosenbach 2005: 617). For each Landmark in the dataset it is established whether it is a ‘simple lexeme’ (e.g. *laud* ‘table’) or a ‘compound’ (e.g. *kirjutuslaud* ‘writing desk’). Similarly to the above argumentation about length of the Landmark phrase, morphological complexity of Landmark is taken as a proxy for general complexity and based on Mondorf’s (2003: 205) analytic support, it is predicted that the analytic, i.e. more explicit, construction is used in cognitively more demanding environments.

Word class of Trajector. Different expression types have been found to affect the choice of syntactic alternatives; see, for example, Bresnan and Ford (2010) for an overview of how this variable affects the dative alternation in English and Gries (1999) for the English particle placement. The word class of Trajector has three levels: ‘noun’, ‘pronoun’, and ‘verb phrase’.

Relative position between Trajector and Landmark. Estonian is considered a language with a relatively free word order (Lindström 2005: 10) and in principle, both constructions can come at the beginning, in the middle, or at the end of a clause. Based on the principle of end-weight which states that “long, complex phrases tend to come at the ends of clauses” (Wasow 1997: 81), we assume that it is the analytic *peal* construction that creates a heavier constituent because it has the extra lexeme (*peal* ‘on’) and should thus prefer the clause-final position. Data are coded for the relative order of Trajector and Landmark phrases. The Landmark phrase either follows (coded as ‘tr_lm’) or precedes (coded as ‘lm_tr’) the Trajector phrase.

Dialect. 10 areal varieties are traditionally identified for the Estonian language. These varieties are further grouped into northern and southern dialects based mostly on phonological, morphological and lexical traits. The northern group is formed by Insular, Western, Mid, Coastal, Northeastern and Eastern dialects, whereas Mulgi, Tartu, Võru and Seto dialects represent the southern group. Sometimes Coastal and Northeastern dialects are distinguished as a separate subgroup within northern dialects. Dialect borders have never been clear-cut and since Estonian dialects have levelled considerably during the past 100 years or so, it is generally more reasonable to talk about preference patterns across dialect continuum in the context of morphosyntactic phenomena, rather than just stating the existence or non-existence of certain features in certain areas. For the purpose of the present study, we have operationalized the dialect variable through the 10 traditional areas that are used in the CED (Eastern, Coastal, Insular, Mid, Mulgi, Northeastern, Seto, Tartu, Võru and Western). Figure 1 presents the map of Estonian dialects and Figure 2 illustrates the data collection points included in the present study.

Figure 1. Estonian dialects[5]

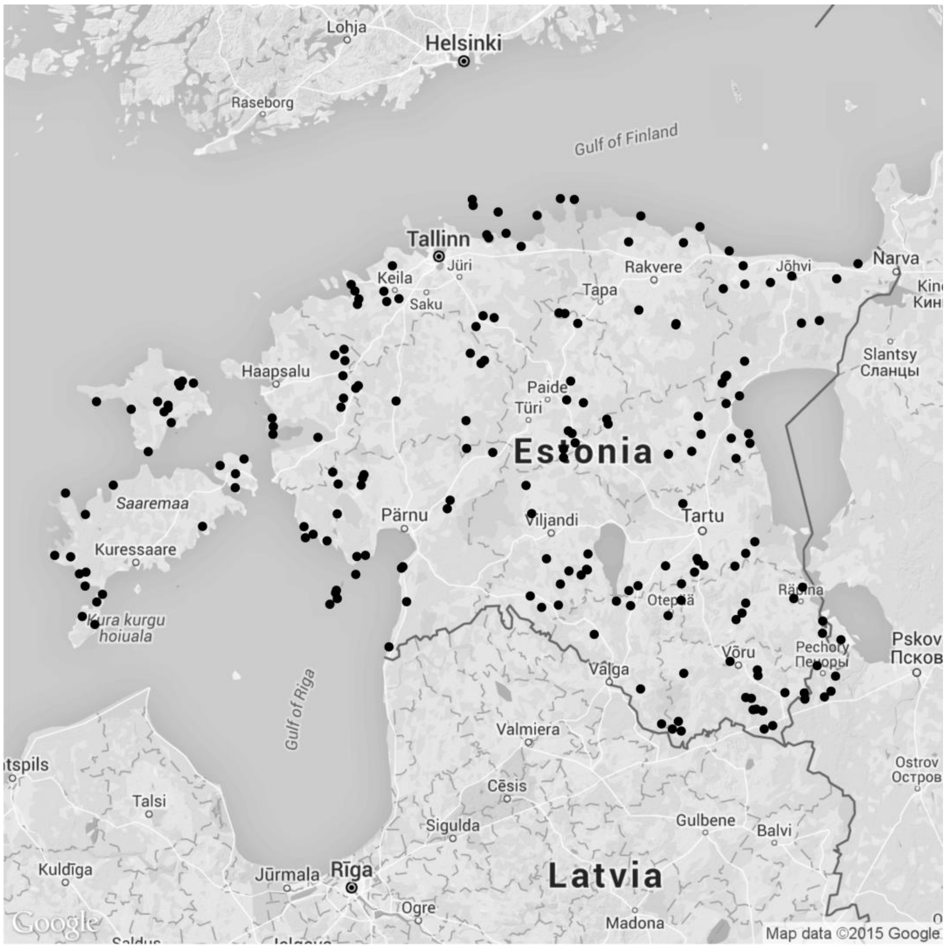


Informant. It has been shown in previous studies that subjects, in our case informants, tend to have their own specific preferences or dispreferences regarding the choice between alternative constructions (e.g. Dąbrowska 2008, 2010; Tagliamonte & Baayen 2012). It is therefore important to introduce the variable informant into the analysis as an additional source of

[5] The dialect map is a modified version of EKI (2014) map.

variation that might influence the structure we are trying to find in the data. In our study, 286 individual speakers are included in the analysis. In the logistic model, it is treated as a random effect.

Figure 2. Data collection points in the present study (Google Maps as a base layer, *ggmap* package (Kahle & Wickham 2013) in R applied for visualization)



Lemma. In a similar fashion to individual speakers, it has been shown that specific words have also their own preference or dispreference for one alternative construction over another (e.g. Stefanowitsch & Gries 2003; Bresnan et al. 2007; Janda, Nesset & Baayen 2010). It is therefore necessary to account for this variability in our statistical analysis as well. The lemmas, however, cannot be used as a variable in the 'tree and forest' analysis, since, trivially enough, a predictor with this many levels requires more computational power than is currently at our disposal. There are 420 different lemmas used with the two constructions in our dataset.

Table 1 is an overview of the dataset: the first column gives the name of the predictor and the second column specifies the corresponding levels.

Table 1. Overview of the dataset (N = 2,032: adessive = 722, *peal* = 1,310)

Predictor	**Levels**
MOBILITY (mobility of Landmark)	mobile, static
TYPE (type of Landmark)	place, thing
VERBLEMMA (lemma of the verb used together with the adessive or *peal construction*)	191 verb lemmas
VERBGROUP	activity, existence, motion, no verb, posture
LENGTH (length of the Landmark phrase in syllables)	ranging from 1 to 9 syllables
COMPLEXITY (morphological complexity of the word used in the adessive or *peal construction*)	compound, simple
TRWORDCLASS (word class of the Trajector phrase)	NP, pronoun, VP
POSITION (relative position between the Trajector and Landmark phrase)	lm_tr, tr_lm
DIALECT (the dialect area)	Coastal, Eastern, Insular, Mid, Mulgi, Northeastern, Seto, Tartu, Võru, Western
INFORMANT	286 informants
LEMMA (lemma of the word used in the adessive or *peal construction*)	420 lemmas

2.3 Methods for statistical analysis of the data

In order to have a better picture of the structure of the dataset and to test the relevance of the predictors described in the previous section, we apply several methods for statistical analysis. We focus on multivariate analysis and use logistic modeling (Harrell 2001; Pinheiro & Bates 2002; Hosmer, Lemeshow & Sturdivant 2013) as well as classification trees and random forests (the ‘tree & forest’ method; Breiman, Friedman, Olshen & Stone 1984; Breiman 2001; Strobl, Boulesteix, Zeileis & Hothorn 2007; Strobl et al. 2009). We use the ‘tree & forest’ method because of its advantages over logistic regression (Baayen et al. 2013: 258): the results are more intuitive, it arrives at its optimal solution on its own (eliminating thus the daunting step of finding the “best” regression model), it does not impose distributional constraints on the dataset, and its output is a classification tree with an intuitive diagram of the outcomes predicted by various combinations of predictor values. However, since our dataset has two variables that have a large number of levels (informant and lemma) and since the ‘tree & forest’ method “given current implementations and hardware limitations, does not scale up to data with many subjects and many items” (Baayen et al. 2013: 277), mixed-effects logistic regression (Pinheiro & Bates 2002; Baayen et al. 2008) will be used to distinguish between variability tied to specific informants and word lemmas and variability attributed to the linguistic predictors of primary interest. Informant will be included in drawing classification trees and for building a random forest model, but variable importances for the latter will not be calculated. Lemma is left out from the ‘tree & forest’ models completely. All of the statistical analyses are carried out using R (version 3.1.2, R development core team 2014), the open-source language and environment for statistical computing, freely available at http://www.r-project.org.

3. Logistic regression

The goal of a logistic regression model is to predict the probability that a given response value will be used by means of the logarithm of the odds ratio. We use mixed-effects logistic regression (Baayen et al. 2008; Pinhero & Bates 2002) to model the ‘response’, i.e. the choice between the adessive case and *peal construction*, as depending on the explanatory variables or predictors described in Section 2.2. The analysis is carried out in R using the function glmer() from the package ‘lme4’ (Bates 2014; Bates, Maechler, Bolker, Walker, Christensen, Singmann & Dai 2015). Our initial aim is to refit the original binary logistic regression model presented in Klavan (2012) to the adessive and *peal* alternation in the database of Estonian dialects. We are interested in finding out whether the same predictors that were found significant for the written, standard language also help to explain (some of) the variation in the spoken, non-standard language. To this end, a binary logistic model was fit to the dialect data described by the following formula:

> CONSTRUCTION ~ LENGTHLOG + COMPLEXITY + MOBILITY + VERBGROUP + TRWORDCLASS + POSITION

All of these six predictors were deemed significant in the model fitted to the standard written language (Klavan 2012: 176–181). The written language model had a fairly conservative fit – the value of *C*[6] was 0.761, just under the critical value of 0.8. When we refit the same model to dialect data, length and position fail to make significant contributions to the model fit. Eliminating these predictors from the model formula results in a model where all the remaining predictors are significant, but the *C* value is low, 0.708. Moreover, fitting the above specified simple binary logistic regression model to the dialect data fails to take into account three crucial points. First, the dialect data comes from 286 informants, majority of whom contribute more than one observation. Second, we also coded the predictor lemma which we assume to play a role and which has as many as 420 different levels. These two predictors are treated as random-effect factors – they are sampled from larger populations of informants and lemmas. Third, we have data from different dialect areas. Dialect is modelled in as a fixed-effect factor – it has ten levels which exhaustively describe all the possible Estonian dialect areas.

3.1 Mixed-effects logistic regression model of Estonian dialect data

Following Baayen et al. (2013) we use a hypothesis-driven search for the best, i.e. the simplest yet most adequate model for the data. We start with a model including all seven predictors (length, complexity, type, verb group, Trajector word class, position, and dialect) modelled as fixed-effects and lemma and informant as random-effects. We remove the insignificant predictors one at a time. This step is carried out on the full set of data twice, after which only significant predictors remain. The best mixed-effects logistic model for the adessive and *peal* alternation in Estonian dialect data is described by the following formula:

> CONSTRUCTION ~ LENGTHLOG + COMPLEXITY + TYPE + VERBGROUP + DIALECT + (1|LEMMA) + (1|INFORMANT)

The mixed-effects logistic regression model yields the estimates for the coefficients shown in Table 2 for the fixed-effect predictors. The first column, ‘Estimate’, gives the estimated coefficients – a logistic model estimates how the log of the odds ratio depends on the predictors. In R, the levels of the response are ordered alphabetically, and it takes the second one of the two alternatives to be a success. For the present data, this means

[6] Here and henceforward the index of concordance, *C*-index (Harrell 2001: 247).

that the *peal* construction is a success, and the model ascertains how the log of the number of *peal* constructions divided by the number of adessive constructions depends on the predictors.

Table 2. Coefficients for a mixed-effects logistic regression model for Estonian dialect dataset

	Estimate	**Std. Error**	**z-value**	***p*-value**
Intercept	-1.894	0.643	-2.947	0.0032
LENGTHLOG	1.390	0.467	2.976	0.0029
COMPLEXITY = simple	1.765	0.442	3.995	0.0001
TYPE = thing	1.379	0.316	4.372	0.0000
VERBGROUP = existence	-0.531	0.177	-2.996	0.0027
VERBGROUP = motion	-1.287	0.234	-5.496	0.0000
VERBGROUP = no verb	-0.142	0.276	-0.515	0.6069
VERBGROUP = posture	-0.180	0.467	-0.386	0.6998
DIALECT = Eastern	1.266	0.550	2.303	0.0213
DIALECT = Coastal	0.270	0.548	0.492	0.6224
DIALECT = Insular	1.137	0.460	2.472	0.0134
DIALECT = Mid	1.663	0.474	3.510	0.0004
DIALECT = Mulgi	1.414	0.590	2.396	0.0166
DIALECT = Seto	2.567	0.754	3.404	0.0007
DIALECT = Tartu	1.919	0.570	3.367	0.0008
DIALECT = Võru	1.665	0.531	3.137	0.0017
DIALECT = Western	2.265	0.477	4.751	0.0000

In R, factors are dealt with by taking one factor level as point of reference. For this particular factor level the group mean is calculated. For the other factor levels, the difference between its group mean and the group mean for the reference level is calculated. All group means are on the logit scale. The first item on the list of coefficients is the Intercept. R chooses as values at the Intercept those that come first alphabetically. Thus the Intercept here represents the group mean (on the logit scale) for the following six predictors: LENGTHLOG = 0.37 (equivalent to mean of LENGTHLOG, the exponential value of which is 1.46, meaning that on average the Landmark phrase is 1.46 syllables long), COMPLEXITY = compound, TYPE = place, VERBGROUP = activity, DIALECT = Northeastern[7]. The

[7] In the final model, we changed the reference level for the predictor dialect from the alphabetically first one to Northeastern. We ran models with all of the 10 values of

negative value of -1.894 tells us that the model predicts the adessive construction here. When we change to another group mean, for LENGTHLOG = 0.37, COMPLEXITY = compound, TYPE = place, VERBGROUP = motion, DIALECT = Northeastern, the group mean is -1.894–1.287 = -3.181, indicating that for observations with a motion verb, the *peal* construction is used even less often.

The second column in Table 2 presents a measure of how uncertain the model is about the estimate for the coefficient. The greater the standard error, the more careful we should be when interpreting the results. The third column, values for the z-scores, is obtained by taking the values in the first column and dividing them by the values in the second column. The final column lists the associated *p*-values which "evaluate how surprised we should be to observe a coefficient with as large (or as small, when negative) a value as actually observed" (Baayen et al. 2013: 263). For the Intercept, the small *p*-value indicates that the group mean for LENGTHLOG = 0.37, COMPLEXITY = compound, TYPE = place, VERBGROUP = activity, DIALECT = Northeastern has a log odds that is significantly below 0. This means that the proportion of the *peal* construction is significantly below 50%. For the other terms with small *p*-values, we have good evidence that the differences in group means are significant.

When using *p*-values to evaluate the coefficients for verb group and dialect for which we compare multiple coefficients (these predictors have more than two levels), we need to apply a method of multiple comparisons. We have opted for the conservative Bonferroni correction according to which the *p*-values are multiplied by the number of comparisons (Crawley 2007: 486). In the present case, we have 4 coefficients for verb group (we are making 4 comparisons), which means that the *p*-values for the coefficients of this predictor need to be multiplied by 4. For dialect, the *p*-values of the coefficients need to be multiplied by 9.

3.2 Performance of the mixed-effects model

We use Akaike's Information Criterion (AIC) and *C* measure to assess the performance of the model. AIC is a statistical measure used to compare models with different numbers of parameters (Hosmer et al. 2013: 120). It

dialect as the reference level. From the perspective of our analysis, having Northeastern as the base gives us a model which reflects the structure of the dataset best.

tells us how close the fitted values of a model tend to be to the true values, in terms of a certain expected value (Agresti 2002: 216). There is no statistical test to compare values of AIC. In general, lower values of AIC are taken to indicate a better model fit. The *C* measure, or the index of concordance a.k.a. the area under the (receiver operating characteristic) curve, provides a description of the fitted model's classification accuracy. It ranges from 0.5 to 1.0 and reflects the model's ability to discriminate between the two outcomes. The following general guidelines are given as a rule of thumb: $C = 0.5$ – no discrimination, like flipping a coin; $0.5 < C < 0.7$ – poor discrimination, not much better than flipping a coin; $0.7 \leq C < 0.8$ – acceptable discrimination; $0.8 \leq C < 0.9$ – excellent discrimination; $C \geq 0.9$ – outstanding discrimination (Hosmer et al. 2013: 177).

Following Baayen et al. (2013: 278) we assess the importance of each predictor in the model by comparing the decrease in AIC as the different predictors are added to the model specification. Table 3 lists the statistics for the decrease in AIC. The first row compares the AIC of a model with informant to a model with only an intercept term. The large decrease in AIC for INFORMANT (86.6) and LEMMA (465.4) indicate that these are the two most important predictors. As also attested in other variationist studies that employ mixed-effects modelling (e.g. Bresnan & Ford 2010: 189; Janda et al. 2010: 40; Baayen et al. 2013: 278), the contributions of the linguistic predictors compared to the random effect terms are much smaller. It can be seen that the most important linguistic predictors are verb group, dialect and type of Landmark. Length also contributes to the model fit, but its contribution is smaller compared to the other predictors. The column 'logLik' lists the model's log likelihood, 'Chisq' is twice the difference in logLik, which follows a chi-squared distribution with as degrees of freedom the number of additional parameters used by the more complex model (column 'Chi.Df') (Baayen et al. 2013: 278). The *p*-value is derived from these chi-squared statistics.

Table 3. Model comparison statistics for the Estonian dialect dataset

	logLik	**Chisq**	**Chi.Df**	***p*-value**	**Reduction in AIC**
INFORMANT	-1277.8				86.6
LEMMA	-1044.2	467.34	1	0.000	465.4
LENGTHLOG	-1041.3	5.7657	1	0.0163	3.7
COMPLEXITY	-1030.2	22.113	1	0.000	20.2
TYPE	-1020.4	19.689	1	0.000	17.6
VERBGROUP	-1001	38.842	4	0.000	30.9
DIALECT	-979.45	43.025	9	0.000	25

The *C* value of 0.94 tells us that the fit of the model is excellent. The accuracy of the model is 87%, where we judge the model to make a correct prediction if the estimated probability for the *peal construction* is greater than or equal to 0.5 and the *peal* construction was actually observed. When always guessing the most frequent alternative, i.e. the *peal* construction, the prediction accuracy would be 64%. In order to test for any significant interactions between the predictors and to gain a more intuitive understanding of the structure in our data, we will now turn to the 'tree & forest' method to analyse the adessive and *peal* alternation in non-standard, spoken language.

4. Classification trees and random forests

4.1 Classification trees

While logistic regression models are parametric, the 'tree & forest' models are non-parametric methods, meaning they do not make any assumptions about the probability distributions of the variables and as such are suitable for dealing with natural language data which is often unbalanced and "messy". The 'tree & forest' methods can also handle the so-called "large *p* small *n*" (small number of observations per predictor (levels)) problems (Strobl et al. 2009) and adequately analyse highly correlated variables (Strobl, Boulesteix, Kneib, Augustin & Zeileis 2008), both of which are characteristic of natural language data.

The 'tree & forest' method makes use of recursive partitioning of the data yielding an optimal sorting of observations with respect to the response variable. For a single classification tree, the observations are divided into binary nodes based on the strength of association between the respective splitting variable and the response variable. Observations with similar response values are grouped together. For each next binary split, the predictor that is most strongly associated with the response variable is selected. Splitting continues until a certain stop condition is reached (Strobl et al. 2009: 325–327). In the current study, *p*-values are employed to select predictors and stop the splitting (for common stop criteria, see Strobl et al. 2009: 327). Once no more significant splits can be made, the partitioning ends and values of the response variable are predicted in each terminal node. As the 'tree & forest' method uses bootstrap sampling (drawing a same-size random sample from a dataset with replacement), it automatically provides a mechanism for validating the model. The sampled data points (the "in-bag" observations) constitute the training dataset for building the model and the data points not sampled (the "out-of bag" observations) are left for testing the predictions made by the model (Baayen et al. 2013: 260). Also, since the predictors that do not contribute to the model are not included in the tree, there is no need to struggle with finding the best model for the data, and because the variations in the sample are random, it is not necessary to specify interactions between predictor variables. In fact, it is even stated that classification trees are best suited for representing complex interactions, while being extremely unlikely to depict only main effects. (Strobl et al. 2009: 329–330.)

Single classification trees are highly intuitive and easily interpretable. The first split partitions the entire sample, the second split partitions only those observations that were grouped by the first split etc. One has to keep in mind, though, that the split selection process here is only considering the previous but not the following splits and as such is only locally optimal (Strobl et al. 2009). In this respect, single classification trees are similar to parametric regression, but it also means that the first split in the tree is one of the most important one, but not necessarily the most important (Strobl et al. 2009: 333; Baayen et al. 2013: 265).

We use the *party* package (Hothorn, Buehlmann, Dudoit, Molinaro & Van Der Laan 2006a; Hothorn, Hornik & Zeileis 2006b; Strobl et al. 2007, 2008) in R to run both the classification tree and random forest analysis on our data. We conducted two separate analyses with slightly different settings. For the first classification tree[8] we excluded the informant variable. In addition to employing *p*-values for the splitting, the stop condition for the terminal nodes is set to 25 to avoid the risk of overfitting, i.e. the minimum number of observations in a terminal node has to be at least 25. If this condition (and adequate significance value) is met, the splitting stops. The classification accuracy of the first tree is 73%. There is an 8% increase in the accuracy of choosing between the adessive or *peal* construction based on the input variables compared to always choosing the most frequent option (the *peal* construction). The index of concordance (*C*) has a value of 0.74, just a bit below the standard threshold 0.8. The results are shown in Figure 3.

[8] set.seed(2000)
ctreeAdePeal = ctree(CX ~ LENGTHLOG + COMPLEXITY + TYPE + VERBGROUP + DIALECT + TRWORDCLASS + POSITION, controls = ctree_control(minbucket = 25), data=dat)

Figure 3. Classification tree without the informant variable

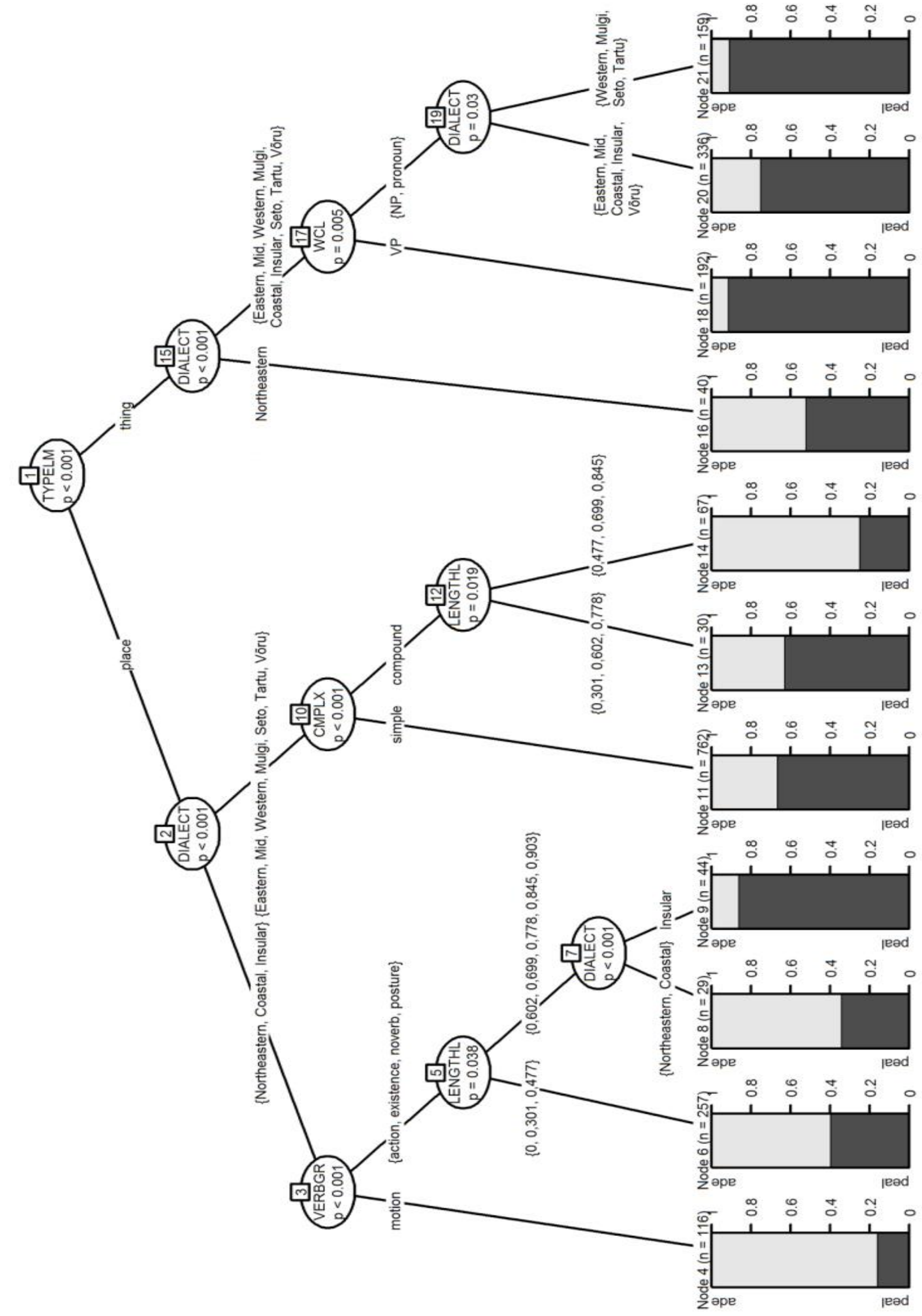


Figure 3 shows that the most significant predictor is the type of Landmark, which splits the data into two groups: places and things. Further splits in both branches are made based on dialects. The left branch of Landmarks which denote places distinguishes Northeastern, Coastal, and Insular dialects from the others. Within this group, the verb class is the next significant predictor. When a motion verb is present, adessive is more commonly used. When it comes to other verbs, length of Landmark introduces additional variation: shorter Landmarks prefer the adessive case, but with longer Landmarks small differences emerge between the dialects. In the Insular dialect *peal* is more frequent with a longer Landmark, while the adessive is more likely to be chosen in the Northeastern and Coastal dialect.

We now turn to the second dialect group which includes the other seven dialects in the place-branch in Figure 3. Here, the next significant split is made at the complexity variable: simple lexemes prefer the *peal construction*, while compounds make more use of the adessive. There seems to be one anomaly in the compound-branch, namely that length splits the compounds into two seemingly incoherent nodes in terms of the short-long distinction: a node with 2-, 4-, and 6-syllable words, and a node with 3-, 5-, and 7-syllable words. When we look at the data, however, half of the 30 cases in the left node are compounds with monosyllabic last components, such as case forms of the lemmas *maan+tee* ('highway') or *heina+maa* ('meadow'), which seem to have a clear preference for *peal* construction and therefore behave more like simple lexemes. In those compounds, the last components' high usage frequency might overshadow the general tendency to use the adessive case with compound nouns[9]. This explanation, however, may be too far-fetched based on such relatively low frequency of these observations and we might as well be dealing with idiosyncratic behaviour of certain lemmas when expressing place in this dialect group. This could be accounted for by including the variable lemma in the 'tree & forest' analysis, given the appropriate technical possibilities, which we currently, unfortunately, lack. We can therefore generalize and say that simple Landmarks tend to use the *peal construction* in this dialect group whereas compounds are more drawn to the adessive case.

For the branch on the right with things as Landmarks in Figure 3, *peal* construction is mostly used. The only exception is the Northeastern dialect that seems to make equal use of both constructions. There appear to be no

[9] We are grateful to the anonymous reviewer for highlighting this point.

other significant predictors that help us to understand the choice between the two constructions, meaning that although we get further significant splits at Trajector word class and dialect, they do not distinguish the response classes on their own, but reflect interactions between different predictor variables.

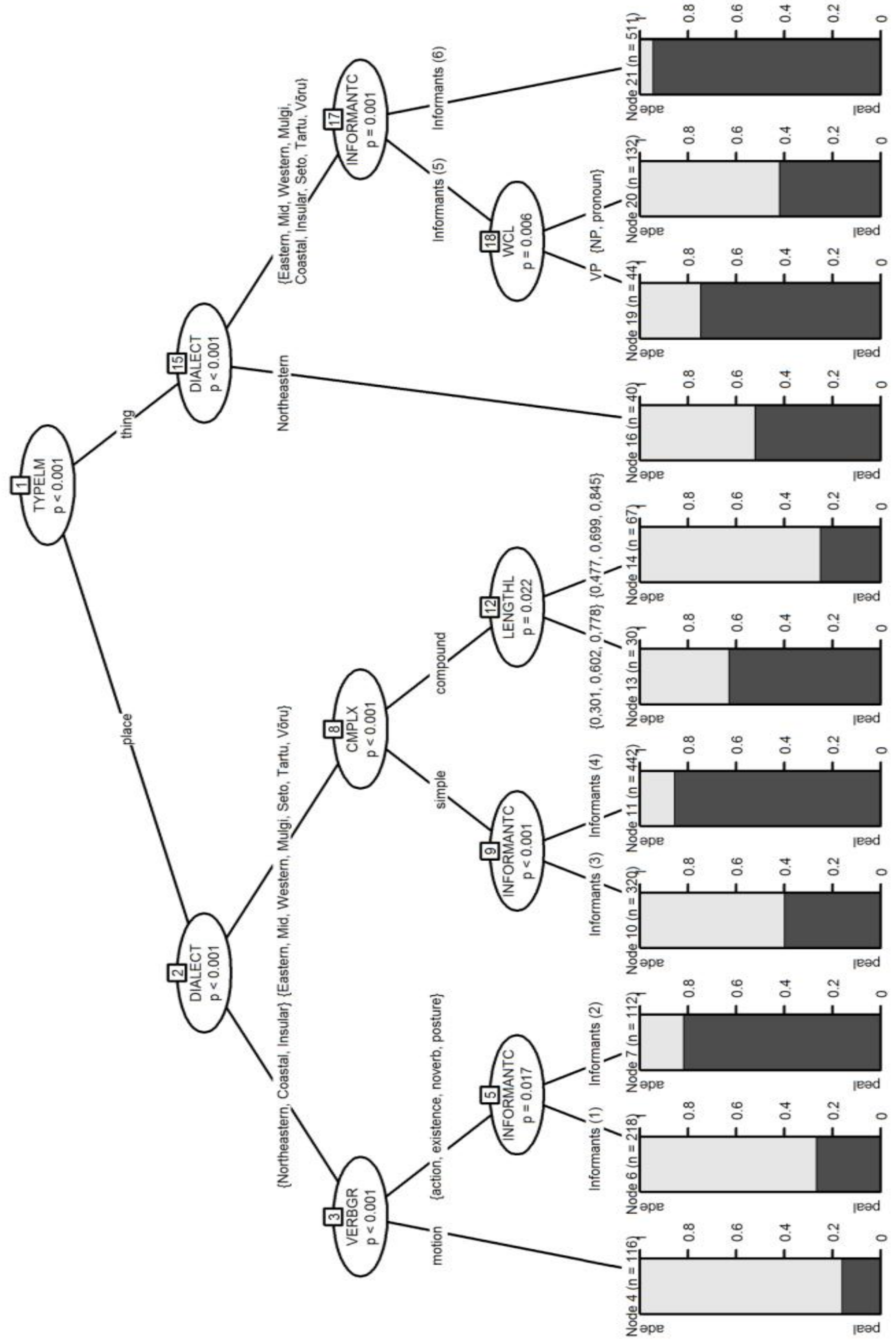
The informant variable was included in the second analysis[10] and the minimum number of observations required in the terminal node was set to 25 again. The classification accuracy increases to 79%; the tree is 14% better than always choosing the more frequent *peal* construction. The *C* value is 0.84. We can therefore conclude that when we include the informant we get a considerably better tree model (Figure 4).

From Figure 4 we can see that the splitting is similar to that in the previous tree in Figure 3. Again the strongest predictor is the type of Landmark. The place-branch is split into two dialect groups as in the the first tree. One group consists of Northeastern, Coastal and Insular dialects. Here the type of verb divides the data further: observations with motion verbs (still preferring the adessive construction) are separated from the rest. However, for all other verbs, the preference for either adessive or the *peal* construction now depends on the speakers and not on dialect or the length of Landmark as in Figure 3. For non-motion verbs in the dialect group of Northeastern, Coastal and Insular dialects, length no longer makes an important distinction once the informant is included. Instead, variation can be attributed to individual preferences.

Length does, however, play a role in the other dialect group (Eastern, Mid, Western, Mulgi, Seto, Tartu, and Võru), where, similarly to Figure 3, compound and hence presumably longer Landmarks have a tendency to occur in adessive constructions, whereas within the group of simple Landmarks, the preference for either construction is now conditioned by the informants.

[10] set.seed(2000)
ctreeAdePealInf = ctree(CX ~ LENGTHLOG + COMPLEXITY + TYPE + VERBGROUP + DIALECT + TRWORDCLASS + POSITION + INFORMANT, controls = ctree_control(minbucket = 25), data=dat)

Figure 4. Classification tree with the informant variable



When we turn to the right branch, where the type of Landmark is ‘thing’, again a distinction is made between Northeastern (making use of both constructions equally) and the other dialects. Compared to the first tree in Figure 3, however, we now observe another variable that introduces a contrast in the response classes in the final nodes – word class of Trajector. The second dialect group is split by informants and the first group of informants is further divided into two nodes distinguishing between verbal and nominal Trajectors: when the Trajector phrase is a verb phrase, the adpositional construction is more common, and when the Trajector phrase is either a noun phrase or a single pronoun, the adessive is more frequently used. There are no further splits in the other group of informants and it is strongly inclined to use the *peal construction*.

While single trees are great for exploring the data to get a sense of its structure and identify possible interactions, they are quite unstable due to the fact that both the cutpoint and the splitting variable selection strongly depend on the structure and distribution of data points on which the model is built. To account for this instability and to draw smoother decision boundaries, an ensemble of classification trees (the ‘forest’) can be aggregated for prediction (Strobl et al. 2009: 330–331).

4.2 Random forests

Random forests introduce yet another source of randomness into the analysis: in addition to randomly sampling the training dataset, it also randomly chooses a set of possible variables to select from for each split, enabling the weaker predictor variables to also affect the final predictions and reveal interactions that would otherwise be missed. That results in a set of diverse single trees that each “vote” for a predicted value for the response variable. The class that gets more votes is chosen as the prediction of the whole forest. Both the number of trees and the number of random variables considered for each split can be set in the model with the arguments *ntree* and *mtry*, respectively. For our analysis, ntree was set to 4000 and mtry to 3. The overall principle in setting these parameters is that the more observations there are in a dataset, the more trees should be generated to guarantee stability, and the more variables are analysed, the larger the set of randomly preselected variables should be. As a default for the latter, a square root of the number of variables for classification and the number of variables divided by three for regression is suggested (Cutler, Edwards, Beard, Cutler, Hess, Gibson & Lawler 2007).

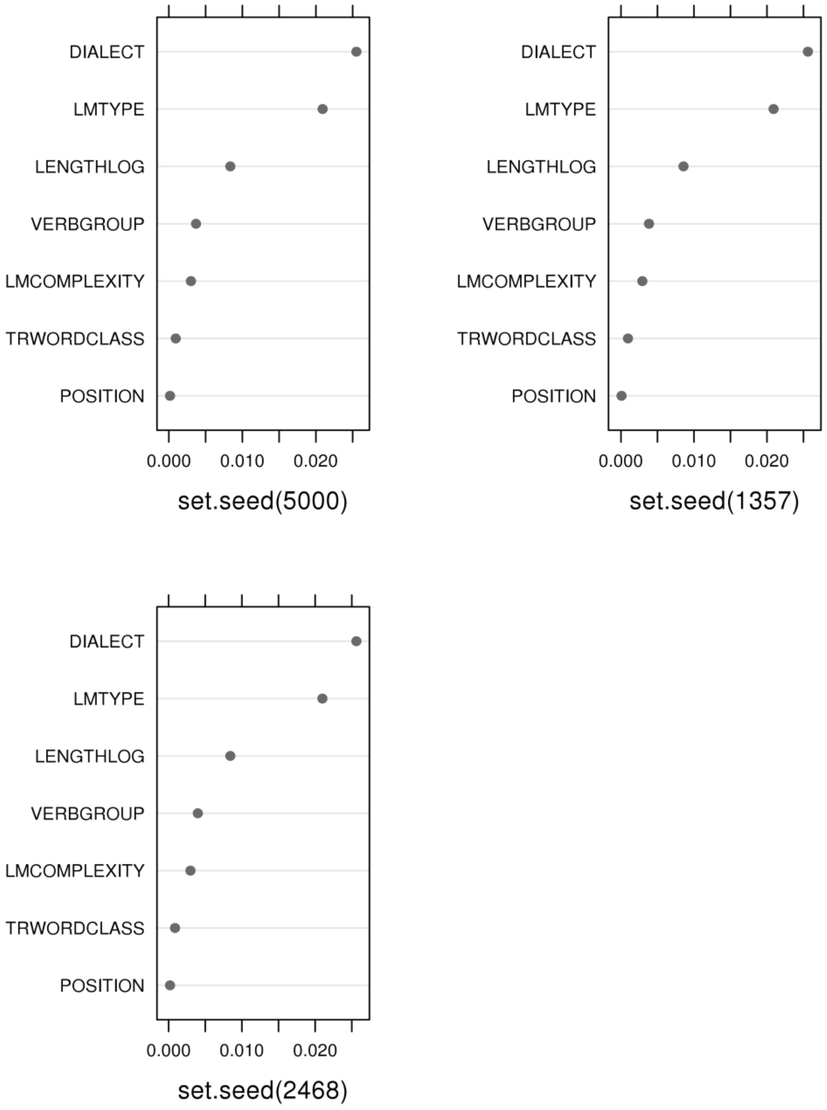
As there is no such thing as an "average tree" to choose for visualization when using ensemble methods, the interpretation of random forests is less straightforward. To evaluate the importance of each predictor across all trees, variable importance measures are computed. The forest therefore only serves us with relative importance of the variables in the model, but does not imply, how the variables affect each other (Baayen et al. 2013: 265–267).

Similarly to the classification tree analyses we also built two random forest models: the first one without the informant variable and the second one with the informant included. Since random forest modelling is computationally extremely demanding, the second analysis (with the informant) could not be completed, since for calculating variable importance measures, even more computational power is necessary. As our resources are currently limited in this respect, we are not able to present variable importance measures for the second analysis.

We built three models with the same setting, but different seeding in order to test the stability of the model and sufficiency of data points. The first model[11] (with all seeding options) achieved a classification accuracy of 76% and the *C*-index for that model is 0.83. The first forest model is therefore better than the first single tree model. Figure 5 presents the variable importance measure graph with three seeding options for the random forest models.

[11] adePealinf.cforest = cforest(CX ~ VERBGROUP + LENGTHLOG + DIALECT + TYPE + COMPLEXITY + POSITION + TRWORDCLASS, data=dat, controls=cforest_unbiased(ntree=4000, mtry=3))

Figure 5. Variable importance in random forest analysis with three different seedings



All the variables that remain on the right from the zero point in Figure 5 influence the choice between the two alternating constructions. The most important predictor is dialect, followed by the type and length of Landmark, significant contributions are also made by verb group, complexity of Landmark, and word class of Trajector. The relative position of Trajector and Landmark does not play a role in distinguishing between the two constructions. The results of the random forest model are similar to the simple trees. Dialect and the type of Landmark play a crucial role in deciding whether the adessive or adpositional construction is chosen to express the locative meaning. The main difference between the forest model and the single trees is that in the former, dialect is more important than the type of Landmark in predicting the response class. Overall, both the single trees (cf. Figures 3 and 4) and the random forest model (cf. Figure 5) indicate that dialect and type of Landmark are by far the two most important predictors compared to the other variables for predicting the choice between the adessive and *peal* construction. Interestingly enough, the forest model without the informant predictor performs worse in accuracy than the second single tree model in which the informant was included. Because of that, the second forest model[12] was ran with similar settings, but with the informant variable included. The overall prediction accuracy rose to 87% and C-index achieved a considerably higher value of 0.95, which gives us similar results as the mixed-effects logistic regression in terms of classification accuracy and model fit. The variable importance measures are not calculated for this model due to the limitations mentioned above.

5. Discussion of results

In general, both the mixed-effects logistic regression and the ‘tree & forest’ method show similar results. Both methods single out dialect, type of Landmark, verb group, complexity and length of Landmark as significant predictors for determining the choice between the adessive and *peal* construction in the spoken dialect data. The mixed-effects model confirms the importance of these linguistic predictors even when we factor in variation tied to individual speakers and lemmas. Word class of the Trajector phrase and the relative position of Trajector and Landmark did not make significant contributions to the mixed-effects logistic regression model. In the mixed-effects model, only the main effects are factored in. The ‘tree & forest’ method, however, highlights a vital characteristic of the structure of the dialect data – there are subtle, but important interactions among the predictor variables. We will now take a look at the results in more detail. We will compare the models fitted to the dialect data to the model fitted to the written language data and focus on the linguistic interpretation of the models.

[12] adePealinf.cforest = cforest(CX ~ VERBGROUP + LENGTHLOG + DIALECT + TYPE + COMPLEXITY + POSITION + TRWORDCLASS + INFORMANT, data=dat, controls=cforest_unbiased(ntree=4000, mtry=3))

A direct and detailed comparison between the binary logistic regression model fitted to the standard, written language data and the mixed-effects logistic regression model as well as the tree-based model fitted to the non-standard, spoken language should be approached with care. The fit of a model (including the significance of the predictors and prediction accuracy) depends entirely on the combination of the specific parameters included in the models – the models fitted to the two datasets differ in model specifications. Our aim is not so much in comparing the performance of the alternative models against each other, but rather to arrive at a best possible description of the structure of the data. Taking a look at some of the trends that emerge from the different models enables us to have a better picture of the alternation between the synthetic adessive construction and the analytic *peal* construction across registers and language varieties.

When we compare the written language model to non-standard, spoken language model a number of differences can be seen. One of the differences pertains to the variable length. Length was one of the strongest predictors in the binary logistic regression model fitted to the standard, written language. Klavan (2012: 188–190) showed that in written language longer Landmark phrases predict the adessive construction, while shorter Landmark phrases predict the *peal* construction (all other things being equal). Length also contributes to the fit of the dialect data, but its significance compared to the other predictors is much lower. Although the mixed-effects logistic regression model shows that in non-standard spoken language longer phrases predict the *peal* construction, we can see from the classification trees that this is not necessarily true. Length is significant because it interacts with dialect and complexity. It is therefore hard to make assumptions about the main effect of length across different registers on the basis of our data.

Another difference between the written language model and the models fitted to dialect data concerns the type of verb used together with the two constructions. In the written language data, verb group makes a relatively low contribution to model fit, while for spoken, dialect data, it makes a relatively large contribution, at least to the fit of the mixed-effects model. In the random forest model, verb ranks relatively low in the ranking of predictors, but single classification trees demonstrate that there are significant interactions between verb group and other predictors, most importantly between verb and dialect. In Northeastern, Coastal and Insular dialects, motion verbs (e.g. *liikuma* ‘to move’) are used with adessive constructions, whereas in other dialects, individual preferences of the speakers (or the length of Landmark in the simple model) affect the choice between the two locative constructions.

Despite the differences, the models fitted to the standard, written language and the non-standard, spoken language also exhibit important similarities. Morphological complexity and type of Landmark (the latter is closely related to the predictor mobility in the written language model, see above Section 2.2) play a similar role in both datasets. In general, compound nouns (e.g. *kirjutuslaud* 'writing-table') prefer the adessive construction. The adessive construction is also preferred when the Landmark is a static place (e.g. *turg* 'market') rather than a mobile thing (e.g. *kapp* 'wardrobe'). Classification trees also imply similar associations. Things are mostly used with *peal* constructions (for certain informants, even more so when the Trajector phrase is a verb phrase), while places are subject to more complex variation: although Northeastern, Coastal and Insular dialects clearly prefer the adessive case with places, in other dialects, complexity of the Landmark is also affecting the choice. Our study therefore confirms the results of Bartens (1978) and Ojutkangas (2008) who detected similar tendencies in the alternation between interior locative cases and the corresponding adpositions in the Saami and Finnish languages respectively: places lend themselves more easily for abstraction and hence are more likely to be used with the adessive, while small manipulable (or movable) objects or things prefer the adposition *peal* since adpositions are more concrete and specific than cases and they convey the meaning of spatial location of an object more clearly (cf. also Comrie 1986; Hagège 2010: 37–38; Lestrade 2010).

As to the regional variation attested in our data, the results of logistic regression show that the use of the synthetic adessive case construction is more frequent in Northeastern and Coastal dialect areas compared to the Mid, Seto, Tartu, Võru and Western dialect areas. The strong preference for adessive constructions in Coastal and especially Northeastern dialects is also evident in the classification trees. This is largely in accordance with Palmeos (1985: 15), since based on the logistic regression analysis, speakers of three out of the four southern dialects (Seto, Tartu, and Võru) as well as the informants from the Western dialect area are more prone to use the analytic adpositional construction, whereas speakers of the northern variants (especially Northeastern and Coastal dialects, which have had closer contacts with Finnish, which is considered to be more conservative in terms of analytic developments) are more likely to make use of the adessive case for expressing locative relationships.

It is difficult to attribute the grouping of the Western and southern dialects to language contacts, but geographically this area clearly forms a "belt" from South-East to North-West. Another interesting fact that our results show is the very clear grouping of Northeastern and Coastal dialects – both clearly prefer the adessive instead of the adpositional construction. This group is also very homogeneous in terms of interindividual variation. We conclude that based on the nominal categories we have looked at, the geographical spread of these analytic and synthetic constructions does not correlate with the distribution of the verbal constructions presented in previous studies (e.g. Serebrennikov 1959; Uiboaed 2013) which leads us to believe that the forces that drive the analytic and synthetic developments are not necessarily the same in verbal and nominal paradigms. For instance, the language contact explanations commonly provided to shed light on the analytical tendencies in Estonian do not apply for the adessive and *peal* alternation.

Although we conclude that the mixed-effects logistic regression model as well as the tree-based models reported in this paper have an excellent fit and a very high prediction accuracy, there are a number of potentially relevant predictors that are missing from our models, but which are shown to make significant contributions to models fitted to other alternation phenomena. Most importantly, frequency and priming effects in all their guises (e.g. semantic, lexical, morphological, syntactic) should be taken into account in the future when modelling both written language data and non-standard, spoken language data. For example, Szmrecsanyi (2005, 2006), Hinrichs and Szmrecsanyi (2007), Bresnan and 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 and factoring in this predictor increases the researcher's ability to account for linguistic variation. There is ample evidence across different studies which demonstrate the importance of priming at all levels of language (cf. references in Szmercsanyi 2005). The effects of priming are especially prominent when we look at conversational interactions as exhibited by the recorded texts in the Corpus of Estonian Dialects.

Other discourse related factors, e.g. text frequency of the lemma (Hinrichs & Szmrecsanyi 2007: 451) and definiteness (cf. studies cited in Bresnan & Ford 2010: 174), have also been shown to play a role across alternating pairs. As a possible hypothesis for the future, it can be predicted that more frequent lemmas prefer the adessive construction, while less frequent lemmas prefer the *peal* construction (cf. Ojutkangas 2008 who demonstrated that in Finnish dialects interior locative cases express conventional spatial relations, while the corresponding adpositional constructions are used when this relation is somewhat unconventional or unexpected for the speaker). Including these and other potentially relevant factors (e.g. phonological and pragmatic) remains an undertaking for the future. At this stage we were mainly interested to see if the factors found significant for the written data also play a role in the non-standard, spoken language and whether there are any differences between dialects.

In our study we proceed from the assumption that corpus-based models allow for a cognitively realistic language description, but this should not be taken for granted. As a potential avenue for future research, a comparison of the corpus-based models to behavioural data (i.e. native speaker performance in psycholinguistic experiments) is called for. Conducting psycholinguistic experiments with Estonian dialect speakers is, however, a complicated task in terms of modern dialect communities. The recordings that the texts in the CED are based on were mostly made in the 1960s and 1970s with non-mobile older rural speakers. Since then (and even already during that time), Estonian dialects have levelled remarkably due to multitude of factors, such as increased mobility and educational opportunities of the speakers, and both common and written language have had a strong influence on the local varieties. As a result of that, only a minority of the local varieties which are distinguishable by prosodic, phonological, lexical, and morphosyntactic features are still alive to a certain extent. Finding the comparable “population” for behavioral data as our CED data is based on would therefore be extremely difficult. We could, however, test whether the same or a different set of predictors is influencing the choice between the two constructions in present-day standard language. This would allow us to assess the cognitive plausibility of our statistical models (cf. Divjak, Arppe & Dąbrowska 2016). Yet another fruitful avenue for future work includes multimodel inference (Burnham & Anderson 2002; Barth & Kapatsinski 2014). The idea is that corpus data often comes with high model selection uncertainty, which is mainly due to the high redundancy evident in language, i.e. every feature is predictable from multiple other features. Collinearity among the possible predictors makes the selection of one, single model problematic and multimodel inference avoids committing to a single model. After all, “all models are wrong” (Box 1976: 792) and “the correct model can never be known with certainty” (Crawley 2007: 339).

6. Conclusion

In this paper, we have shown that the alternation between the synthetic adessive construction and the analytic *peal* 'on' construction is not free and we have identified a number of semantic and morpho-syntactic factors that influence the choice between the two constructions in non-standard, spoken language. We have presented multivariate statistical analyses and constructed models that can predict the choice between alternative constructions with a 94% of classification accuracy. Although we have seen that the specific lemma (and to a lesser extent the specific informant) contributes significantly to predicting the choice between the two alternatives, mixed-effects logistic regression also confirmed the importance of the linguistic fixed-effects. Both logistic regression and the 'tree & forest' method confirm that 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*. In addition, the 'tree & forest' analysis demonstrates important interactions among the predictor variables. We have shown that the predictors type and complexity of the Landmark hold across register – they play a role in both the standard, written language as well as in the non-standard, spoken language. Overall, our study confirms the results of the previous studies on the alternation between locative cases and the corresponding adpositions (Bartens 1978; Ojutkangas 2008; Klavan 2012). Most importantly, places are more likely to be used with the adessive, while small manipulable (or movable) objects or things prefer the adposition *peal*.

In addition, we have demonstrated that there are significant differences between Estonian dialects as pertains to the use of these alternative constructions. Our results confirm the observations made by Palmeos (1985: 15) that the Western and southern dialects tend to choose the analytic *peal* construction instead of the adessive in the same linguistic context. Another important result that our study shows is the very clear grouping of Northeastern and Coastal dialects – both clearly prefer the adessive instead of the adpositional construction.

We have also shown the benefits of methodological pluralism – we have applied both mixed-effects logistic regression as well as the 'tree & forest' method to analyse one and the same linguistic data. Both methods have their own advantages (as well as disadvantages) and applying them in tandem gives us a better picture of the linguistic data we are studying. On the one hand, the mixed-effects logistic regression enables us to distinguish between variability tied to specific informants and word lemmas and variability attributed to the linguistic predictors of primary interest. The ‘tree & forest’ method, on the other hand, enables us to test for any significant interactions between the various predictors and it outputs a diagram which is much more intuitive than the output of a mixed-effects logistic regression model. The bonus of using both methods is that we have extra confidence in the reliability of our results – the two alternative constructions show different patterns in terms of variation in meaning, register and dialect.

References

Agresti, Alan (2002) *Categorical Data Analysis*. New York: Wiley-Interscience.

Anttila, Arto; Adams, Matthew & Speriosu, Michael (2010) The role of prosody in the English dative alternation. *Language and Cognitive Processes* 25 (7–9): 946–981.

Arnold, Jennifer E.; Wasow, Thomas; Losongco, Anthony & Ginstrom, Ryan (2000) Heaviness vs. newness: the effects of complexity and information structure on constituent ordering. *Language* 76 (1): 28–55.

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

Baayen, R. Harald; Davidson, Douglas J. & Bates, Douglas M. (2008) Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59 (4): 390–412.

Barth, Danielle & Kapatsinski, Vsevolod (2014; ahead of print) A multimodel inference approach to categorical variant choice: construction, priming and frequency effects on the choice between full and contracted forms of *am, are* and *is*. *Corpus Linguistics & Linguistic Theory*: 1–58.

Bartens, Raija. 1978. *Synteettiset ja analyyttiset rakenteet lapin paikanilmauksissa* [Synthetic and Analytic Constructions in Saami Locative Expressions]. Suomalais-ugrilaisen Seuran toimituksia 166. Helsinki: Suomalais-Ugrilainen Seura.

Bates, Douglas (2014) Computational methods for mixed models. <http://cran.r-project.org/web/packages/lme4/vignettes/Theory.pdf> (27 March 2015).

Bates, Douglas; Maechler, Martin; Bolker, Ben; Walker, Steven; Christensen, Rune Haubo Bojesen; Singmann, Henrik & Dai, Bin (2015) lme4. <http://cran.r-project.org/web/packages/lme4/lme4.pdf> (27 March 2015).

Box, George E. P. (1976) Science and Statistics. *Journal of the American Statistical Association* 71 (356): 791–799.

Breiman, Leo (2001) Random Forests. *Machine Learning* 45 (1): 5–32.

Breiman, Leo; Friedman, Jerome; Olshen, Richard A. & Stone, Charles J. (1984) *Classification and Regression Trees*. Belmont, Calif: Wadsworth.

Bresnan, Joan (2007) Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In Sam Featherston & Wolfgang Sternefeld (eds.), *Roots: Linguistics in Search of Its Evidential Base*, pp. 77–96. Berlin: Mouton de Gruyter.

Bresnan, Joan; Cueni, Anna; Nikitina, Tatiana & Baayen, R. Harald (2007) Predicting the Dative Alternation. In Gerlof Bouma, Irene Krämer & Joost Zwarts (eds.), *Cognitive Foundations of Interpretation*, pp. 69–94. Amsterdam: Royal Netherlands Academy of Science.

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

Burnham, Kenneth P. & Anderson, David R. (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach.* 2nd ed. New York: Springer.

CED (2015) Corpus of Estonian Dialects. <http://www.murre.ut.ee/mkweb>.

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

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

Cutler, Richard D.; Edwards Jr., Thomas C.; Beard, Karen H.; Cutler, Adele; Hess, Kyle T.; Gibson, Jacob & Lawler, Joshua J. (2007) Random forests for classification in ecology. *Ecology* 88: 2783–2792. Appendix A. *Technical details* and *additional capabilities* of *random forests*. <http://www.esapubs.org/archive/ecol/E088/173/appendix-A.htm> (27 March 2015).

Dąbrowska, Ewa (2008) Questions with long distance dependencies: a usage-based perspective. *Cognitive Linguistics* 19 (3): 391–425.

—— (2010) Naive v. expert intuitions: an empirical study of acceptability judgements. *The Linguistic Review* 27 (1): 1–23.

Divjak, Dagmar & Arppe, Antti (2013) Extracting prototypes from exemplars. What can corpus data tell us about concept representation? *Cognitive Linguistics* 24 (2): 221–274.

Divjak, Dagmar; Arppe, Antti & Dąbrowska, Ewa (2016) Machine Meets Man: Evaluating the psychological reality of corpus-based probabilistic models. *Cognitive Linguistics* 27 (1).

EKI (2014) = Eesti Keele Instituudi kohanimeandmebaasi kihelkonnapiiride andmestik [Map of the Place Name Database created by the Institute of Estonian Language].

Erelt, Mati & Metslang, Helle (1998) Oma või võõras? [Our own or foreign?] *Keel ja Kirjandus* 10: 657–668.

Erelt, Mati; Kasik, Reet; Metslang, Helle; Rajandi, Henno; Ross, Kristiina; Saari, Henn; Tael, Kaja & Vare, Silvi (1995) *Eesti keele grammatika I. Morfoloogia* [The Grammar of Estonian I. Morphology]. Tallinn: Eesti Teaduste Akadeemia Eesti Keele Instituut.

Erelt, Mati; Erelt, Tiiu & Ross, Kristiina (2007) *Eesti keele käsiraamat* [Handbook of Estonian]. Tallinn: Eesti Keele Sihtasutus.

Gries, Stefan Th. (1999) Particle movement: a cognitive and functional approach. *Cognitive Linguistics* 10 (2): 105–145.

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.

Hawkins, John A. (1994) *A Performance Theory of Order and Constituency*. Cambridge: Cambridge University Press.

—— (2004) *Efficiency and Complexity in Grammars*. Oxford: Oxford University Press.

Hinrichs, Lars & Szmrecsanyi, Benedikt (2007) Recent changes in the function and frequency of standard English genitive constructions: a multivariate analysis of tagged corpora. *English Language and Linguistics* 11 (3): 437–474.

Hosmer Jr, David W.; Lemeshow, Stanley & Sturdivant, Rodney X. (2013) *Applied Logistic Regression*. New Jersey: John Wiley & Sons.

Hothorn, Torsten; Buehlmann, Peter; Dudoit, Sandrine; Molinaro, Annette & Van Der Laan, Mark (2006a) Survival Ensembles. *Biostatistics* 7 (3): 355–373.

Hothorn, Torsten; Hornik, Kurt & Zeileis, Achim (2006b) Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics* 15 (3): 651–674.

Janda, Laura A.; Nesset, Tore & Baayen, Harald R. (2010) Capturing correlational structure in Russian paradigms: A case study in logistic mixed-effects modeling. *Corpus linguistics and linguistic theory* 6 (1): 29–48.

Kahle, David & Wickham, Hadley (2013) ggmap: A package for spatial visualization with Google Maps and OpenStreetMap. R package version 2.3. <http://CRAN.R-project.org/package=ggmap.>

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.

Langacker, Ronald W. (2008) *Cognitive Grammar. A Basic Introduction*. Oxford: Oxford University Press.

Langemets, Margit; Tiits, Mai; Valdre, Tiia; Veskis, Leidi; Viks, Ülle & Voll, Piret (2009) *Eesti kirjakeele seletussõnaraamat.* 4 P–R [Dictionary of Written Estonian. Vol 4 P–R]. Tallinn: Eesti Keele Sihtasutus.

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

Levin, Beth (1993) *English Verb Classes and Alternations: A Preliminary Investigation.* Chicago: University of Chicago Press.

Lindström, Liina (2005) *Finiitverbi asend lauses. Sõnajärg ja seda mõjutavad tegurid suulises eesti keeles* [The position of the finite verb in a clause: word order and the factors affecting it in Spoken Estonian]. Dissertationes Filologiae Estonicae Universitatis Tartuensis 16. Tartu: Tartu Ülikooli Kirjastus.

McDonald, Janet L.; Bock, Kathryn & Kelly, Michael H. (1993) Word and world order: semantic, phonological and metrical determinants of serial position. *Cognitive Psychology* 25 (2): 188–230.

Metslang, Helle (1997) On the use of the Estonian past tense forms through the last one hundred years. In Mati Erelt (ed.), *Estonian: Typological Studies II*, pp. 98–145. Tartu: Tartu University Press.

—— (2009) Estonian grammar between Finnic and SAE: some comparisons. *STUF* 62 (1/2): 49–71.

Mondorf, Britta (2003) Support for *more*-support. In Günter Rohdenburg & Britta Mondorf (eds.), *Determinants of Grammatical Variation in English*, pp. 251–304. Berlin: Mouton de Gruyter.

Ojutkangas, Krista (2008) Mihin suomessa tarvitaan sisä-grammeja? [When are sisä grams used in Finnish?]. *Virittäjä* 112 (3): 382–400.

Palmeos, Paula (1985) *Eesti keele grammatika II. Kaassõna* [The Grammar of Estonian II. Adposition]. Tartu: TRÜ trükikoda.

Pinheiro, José C. & Bates, Douglas M. (2002) *Mixed-Effects Models in S and S-PLUS*. New York: Springer.

R Development Core Team (2014) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Version 3.1.2 <http://www.R-project.org/>

Rohdenburg, Günter (2003) Cognitive complexity and *horror aequi* as factors determining the use of interrogative clause linkers in English. In Günter Rohdenburgand & Britta Mondorf (eds.), *Determinants of Grammatical Variation in English*, pp. 205–249. Berlin: Mouton de Gruyter.

Rosenbach, Anette (2005) Animacy versus weight as determinants of grammatical variation in English. *Language* 81 (3): 613–644.

Serebrennikov, B.A. (1959) Pluskvamperfekti ja perfekti päritolu probleemist läänemeresoome keeltes [The origin of the plusquamperfect in Finnic languages]. In *Emakeele Seltsi aastaraamat IV (1958),* pp. 249–255. Tallinn: Eesti Riiklik Kirjastus.

Stefanowitsch, Anatol & Gries, Stefan Th. (2003) Collostructions: investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8 (2): 209–243.

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).

Strobl, Carolin; Boulesteix, Anne-Laure; Kneib, Thomas; Augustin, Thomas & Zeileis, Achim (2008) Conditional variable importance for random forests. *BMC bioinformatics* 9 (307).

Strobl, Carolin; Malley, James & Tutz, Gerhard (2009) An introduction to recursive partitioning: rationale, application and characteristics of classification and regression trees, bagging and random forests. *Psychological Methods* 14 (4): 323–348.

Szmrecsanyi, Benedikt (2005) Language users as creatures of habit: a corpus-linguistic analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory* 1 (1): 113–150.

—— (2006) *Morphosyntactic Persistence in Spoken English. A Corpus Study at the Intersection of Variationist Sociolinguistics, Psycholinguistics, and Discourse Analysis. Trends in Linguistics.* Studies and Monographs 177. Berlin and New York: Mouton de Gruyter.

—— (2009) Typological parameters of intralingual variability: grammatical analyticity versus syntheticity in varieties of English. *Language Variation and Change* 21 (3): 319–353.

—— (2012) Analyticity and syntheticity in the history of English. In Terttu Nevalainen & Elizabeth Closs Traugott (eds.), pp. 654–665. *The Oxford Handbook of the History of English*. Oxford: Oxford University Press.

—— (2013) Diachronic Probabilistic Grammar. *English Language and Linguistics* 1 (3): 41–68.

Tagliamonte, Sali A. & Baayen, Harald R. (2012) Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change* 24 (02): 135–178.

Uiboaed, Kristel (2013) Verbiühendid eesti murretes [Verbal constructions in Estonian dialects]. Dissertationes Philologiae Eestonicae Universitatis Tartuensis 34. Tartu: University of Tartu Press.

Uiboaed, Kristel; Hasselblatt, Cornelius; Lindström, Liina; Muischnek, Kadri & Nerbonne, John (2013) Variation of verbal constructions in Estonian dialects. *Literary & Linguistic Computing* 28 (1): 42–62.

Wasow, Thomas (1997) Remarks on grammatical weight. *Language Variation and Change* 9 (1): 81–105.

—— (2002) *Postverbal Behaviour*. CSLI Publications.

Wolk, Christoph; Bresnan, Joan; Rosenbach, Anette & Szmrecsanyi, Benedikt (2013) Dative and genitive variability in Late Modern English: Exploring cross-constructional variation and change. *Diachronica* 30 (3): 382–419.

Contact information:

Jane Klavan
University of Tartu
Institute of Estonian and General Linguistics
Jakobi 2–442
Tartu
Estonia
e-mail: jane(dot)klavan(at)ut(dot)ee

Maarja-Liisa Pilvik
University of Tartu
Institute of Estonian and General Linguistics
Jakobi 2–430
Tartu
Estonia
e-mail: maarja-liisa(dot)pilvik(at)ut(dot)ee

Kristel Uiboaed
University of Tartu
Institute of Estonian and General Linguistics
Jakobi 2–430
Tartu
Estonia
e-mail: kristel(dot)uiboaed(at)ut(dot)ee