ESUKA – JEFUL 2021, 12–1: 153–188

THE ALTERNATION BETWEEN EXTERIOR LOCATIVE CASES AND POSTPOSITIONS IN ESTONIAN WEB TEXTS

Jane Klavan
University of Tartu, EE
jane.klavan@ut.ee

Abstract. A probabilistic grammar approach to language assumes that grammatical knowledge has a probabilistic component and that this probabilistic knowledge of language is derived from language experience. It is assumed that the extent and nature of grammatical knowledge is reflected in language variation. In the present paper, the probabilistic variation patterns of the Estonian exterior locative cases and the corresponding postpositions are determined by exploring a large, manually annotated dataset of Estonian web texts. It is proposed that there are both similarities and differences in the morphosyntactic knowledge on the part of Estonian speakers as pertains to the three alternations: allative ~ *peale* 'onto', adessive ~ *peal* 'on', ablative ~ *pealt* 'off'. The study points towards the stability and direction of the factors that have been found significant in the previous studies. Multivariate analysis of corpus data shows that the grammatical knowledge of Estonian exterior cases and the corresponding postpositions is probabilistic and regulated by both morphosyntactic and semantic factors.

Keywords: locative cases, postpositions, constructional alternatives, language variation, probabilistic grammar, mixed models, Estonian

DOI: https://doi.org/10.12697/jeful.2021.12.1.05

1. Introduction

The present paper takes a probabilistic grammar perspective on the Estonian morphosyntactic alternation between exterior locative cases (adessive, allative, ablative) and the corresponding postpositions (*peal* 'on', *peale* 'onto', *pealt* 'off'). The aim of the study is to explore the probabilistic variation patterns of the locative cases and the corresponding postpositions as attested in Estonian web texts. As such, the study is situated at the crossroads of research on usage-based theoretical linguistics (Bybee & Hopper 2001) and variationist linguistics (Tagliamonte 2011). The aim of the study is to advance our understanding of

the morphosyntactic knowledge on the part of Estonian language users. The study proceeds from the assumption that variation between different ways of saying the same thing is "sensitive to multiple and sometimes competing constraints which influence linguistic choice-making in subtle, probabilistic ways" (Grafmiller et al. 2018). Mixed-effects logistic regression is used on a richly annotated corpus sample of Estonian morphosyntactic alternations to capture the speakers' multivariate and probabilistic knowledge of these alternations quantitatively.

Underlying the probabilistic approach to language are two basic tenets: 1) grammatical knowledge has a probabilistic component, 2) this probabilistic knowledge is derived from language experience (Grafmiller et al. 2018). The aim of this line of research is to measure the extent and nature of grammatical knowledge as it is reflected in language variation. The main focus of the probabilistic grammar framework has been on English syntactic alternations, e.g. the dative alternation (Bresnan 2007, Bresnan & Hay 2008, Bresnan & Ford 2010) and the genitive alternation (Heller, Szmrecsanyi & Grafmiller 2017, Szmrecsanyi 2013, Heller & Szmrecsanyi 2019). The present study makes an important contribution to the canon of studies taking a probabilistic grammar approach by looking at the same set of variables across three alternating pairs in a non-Indo-European language. The multivariate analysis of corpus data shows that the grammatical knowledge of Estonian exterior cases and the corresponding postpositions is probabilistic and regulated by various morphosyntactic and semantic factors, differently from the syntactic alternations in English, where the main constraining factors have been discourse-related factors (e.g. animacy, givenness, weight).

The contributions in Bod, Hay and Jannedy (2003a) indicate that probabilistic mechanisms are characteristic to all levels of language, including morphosyntax. From a sizable body of previous research we know that variation within and across varieties of the same language is very systematic and that this variation is multifactorial and probabilistic (e.g. Bresnan et al. 2007, Bresnan & Hay 2008, Bresnan & Ford 2010, Gries 2003, Röthlisberger, Grafmiller & Szmrecsanyi 2017, Szmrecsanyi et al. 2016). Differently from rule-based approaches to grammar which assume that grammatical knowledge is categorical, and possibly biologically innate, the probabilistic approaches to language assume a "cline of well-formedness" (Bod, Hay & Jannedy 2003b: 4).

Furthermore, grammar is taken to be inherently variable and patterns in variation are thought to be learned from exposure to language use (Bybee & Hopper 2001). Variation itself is assumed to be shaped by social, cognitive or functional factors that influence the production and comprehension of individual speakers. The fundamental differences between rule-based and probabilistic approaches to grammar elicit the use of different methodologies. While rule-based approaches are interested in the categorical (un-)grammaticality of linguistic forms and the preferred method tends to be acceptability judgement tasks, probabilistic approaches use methods that allow to tackle the multivariate, probabilistic nature of language. In this respect, the methodologies and research questions in the probabilistic grammar framework are very much congruent with work in variationist sociolinguistics (e.g. Labov 1972, 1982). Both approaches focus on finding out how and why people choose between alternative ways of saying the same thing.

Convincing evidence to support the probabilistic nature of grammar comes from studies that have explicitly compared corpus-based findings against experimental findings; see Klavan and Divjak (2016) for a survey paper. These studies show that the likelihood of finding a particular linguistic variant in a particular context in a corpus corresponds to the intuition that speakers have about the acceptability or the preferred choice of the variants. Klavan and Veismann (2017), for example, used a forced choice task and an acceptability judgement task based on authentic corpus materials as stimuli to model subjects' responses regarding the naturalness of adessive and *peal* constructions in context. Subjects' responses were compared to the predictions of the regression model fitted by Klavan (2012) to the adessive ~ *peal* alternation. It was shown that subjects' ratings and choices overlapped significantly with corpus-based probability estimates. Similar converging results across corpus-based and experimental studies have been found for the English dative alternation (Bresnan 2007, Bresnan & Ford 2010), the Russian verbs denoting the concept of "try" (Divjak 2010, Divjak & Arppe 2013, Divjak, Arppe & Dąbrowska 2016), and the verbs meaning "come" in Modern Standard Arabic (Abdulrahim 2013, Arppe & Abdulrahim 2013).

It is fairly safe to assume, therefore, that speakers' implicit knowledge about language, not only knowledge of constructional alternatives, is probabilistic in nature. The theoretical approach taken in this study

proceeds from a model of grammar that takes grammar to be the "cognitive organization of one's experience with language" (Bybee 2006: 711). Experience with language is inherently variable – variation within and across languages is highly systematic and conditioned by social, cognitive or functional factors (e.g. Gries 2003, Bresnan & Hay 2008, Klavan & Divjak 2016, Szmrecsanyi et al. 2016). This usage-based approach differs from rule-based approaches that consider linguistic variation as "theoretically irrelevant to the investigation of the principles that determine syntactic structure" (Grafmiller et al. 2018).

Probabilistic variation analysis tends to be based on the analysis of naturalistic corpus data. This is also the methodological approach adopted in the present paper with a focus on present-day Estonian web texts. What I am interested in is the aggregated result of the influences of the cognitive and/or functional factors on individual speakers' language production as attested in population-level linguistic patterns. Proceeding from usage-based approaches (e.g. Bybee & Hopper 2001), it is assumed that individual-level behaviour leads to population-level language patterns and that individual behaviours are guided, to a certain extent, by universal cognitive processes. Three specific predictions can be put forward in this context (cf. Grafmiller et al. 2018): (1) the influence of certain factors on the morphosyntactic variation between exterior locative cases and the corresponding postpositions across different varieties of the Estonian language should be relatively stable in terms of the direction of those factors; (2) the strength of different factors on speakers' choices will vary by the types and frequencies of constructions; (3) the variation in the use of exterior locative cases and the postpositions may be driven by stylistic preferences, situational forces or by cognitive pressures related to language processing.

The present study tackles the prediction about the stability and direction of the various factors across different varieties of Estonian by looking at a variety of Estonian – web texts – that has not been used previously and compares these findings with previous work on Estonian exterior locative cases and postpositions in written (Klavan 2012, 2020) and spoken Estonian (Klavan, Pilvik & Uiboaed 2015). The present study can be therefore seen as a replication of the previous studies. As for the second and third prediction regarding the strength of different factors across different constructions and pinpointing the specific factors that drive the choice between alternating constructions, one way

to test these predictions quantitatively is to use statistical modelling techniques on the corpus data of three alternating pairs that are manually annotated for a specific set of factors. The modelling technique chosen for the present study is mixed-effects logistic regression. By exploring systematically the three predictions highlighted above and by looking at morphosyntactic alternations in a morphologically rich Finno-Ugric language, the present study contributes to the theoretical and empirical discussion of probabilistic grammar analysis by adding to the body of knowledge of the nature and limits of grammatical variation.

2. Alternation between Estonian exterior locative cases and postpositions

The Estonian language exhibits a typologically intriguing language phenomenon – the parallel use between the synthetic locative cases and the analytic postpositions. Estonian reference grammars usually make very general claims on the lines that the meaning of adpositions is more concrete and specific than that of the cases and that the meaning of cases is much more abstract and their range of uses more broader (Erelt et al. 1995: 33–34, Erelt, Erelt & Ross 2007, Erelt et al. 2007: 191, Veismann & Erelt 2017: 446). The studies about other Finno-Ugric languages, however, have provided more specific details. For example, Bartens (1978) shows that in the Saami languages the adpositional constructions are used together with smaller, manipulable things as well as with vehicles. Ojutkangas (2008) specifies that the interior locative cases express conventional spatial relations, while the corresponding adpositional constructions are used when this relation is somewhat unconventional.

The only pair of alternations that has been studied quantitatively using state-of-the-art statistical analysis is the adessive ~ *peal* alternation. For example, Klavan (2012) conducted a comprehensive study that combines a corpus-based study of present-day written Estonian with data from two linguistic experiments. 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.

Similar results have been found for the Estonian adessive ~ *peal* alternation (Klavan 2012, and 2020, Klavan, Pilvik & Uiboaed 2015, Klavan & Veismann 2017). The main morphosyntactic variables that play a role in the choice between the adessive case and the postposition *peal* is the length and complexity of the phrase that is inflected for the case or is used with the postposition[1]: the longer and more complex the phrase is, the more probable it is that the preferred choice is the case ending. The strongest semantic variable that plays a role is the type of entity expressed by the noun: smaller and mobile entities like a chair or a table occur with the postpositional construction and bigger and static entities like a street or a meadow occur with the case construction. Klavan (2012, 2020) has confirmed these results for present-day written Estonian (mainly fiction and newspaper texts) and Klavan, Pilvik and Uiboaed (2015) for non-standard spoken Estonian. Klavan and Veismann (2017) have provided additional experimental evidence to support the corpus-based claims. The aim of this paper is to broaden the scope of previous studies by looking at three alternations simultaneously (thereby extending the scope of alternating pairs) and by extending the source of corpus data from written Estonian to web-based texts (thereby extending the scope of variety). The focus of the study is on a detailed and systematic quantitative, multivariate corpus-based analysis of the alternation between the three locative cases and the corresponding postpositions. Before presenting the analysis itself, a brief typological overview of all six constructions and their main functions is given.

In Estonian, nominals (including nouns and pronouns) are inflected for number and case. Estonian has 14 nominal cases, both in singular and plural; three of them are called "exterior locative cases": allative (all), adessive (ade), ablative (abl). Both the exterior locative cases and the alternating postpositions express spatial relations of an open surface and they form a three-part series expressing direction, location and source respectively (see Table 1).

1 The present paper adopts the terminology from Langacker's (2008: 70) Cognitive Grammar approach 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. In the present study, the entity inflected for the allative, adessive and ablative case is the Landmark phrase, as is the entity inflected for the genitive case followed by the postposition *peale*, *peal* and *pealt*.

Table 1. The system of Estonian exterior cases and the alternating postpositions as exemplified by the noun *kivi* 'stone'.

	LATIVE (direction)	LOCATIVE (location)	SEPARATIVE (source)
Exterior locative cases	*kivile* 'onto the stone'	*kivil* 'on the stone'	*kivilt* 'off the stone'
Alternating postpositions	*kivi peale* 'onto the stone'	*kivi peal* 'on the stone'	*kivi pealt* 'off the stone'

The Estonian locative cases and the postpositions normally take the role of an adverbial (as in *laual / laua peal* 'on the table' in example 1) or adverbial modifier (as *vaas laual / laua peal* 'the vase on the table' in example 2) (Erelt et al. 1995: 58).

(1) *Vaas* *on* *{laual /* *laua* *peal.}*
vase.SG.NOM be.PRS.3SG table.SG.ADE table.SG.GEN on
'The vase is on the table.'

(2) *Vaas* *{laual /* *laua* *peal}* *on*
vase.SG.NOM table.SG.ADE table.SG.GEN on be.PRS.3SG
ilus.
pretty.SG.NOM
'The vase on the table is pretty.'

The Estonian (exterior) locative cases fulfil many functions besides location and many of the functions are relatively abstract. For all three cases in the series, it is more frequent for the case construction to express either temporal relations or addressees, experiencers, possessors, agents, and sources than location. For the adessive case in particular, some linguists have objected to referring to it as a locative case (e.g. Matsumura 1994). The localist theory, however, posits that the concrete uses of a case are more primary than the more abstract uses (Anderson 2006: 95–96, Lyons 1977: 718–724). Even though the raw frequencies of a corpus analysis show that the more abstract uses of the locative cases are much more frequent than the locative uses (see Section 3 of the present paper for one set of such raw frequencies), expressing locations is still an important function of Estonian exterior cases. Following is a list of the different functions carried by Estonian exterior locative cases.

Although not directly the focus of the present study, it is believed that the polysemy of grammatical constructions influences the synonymous relationships these constructions can enter into with other grammatical constructions. It is therefore necessary to be aware that the locative cases have an array of functions.

(3) Functions of the Estonian allative case (adapted from Erelt, Erelt & Ross 2007: 249):
   a. Direction of location: *Mari pani vaasi lauale.* 'Mari put the vase on(to) the table.'
   b. Time: *Koosolek viidi üle neljapäevale.* 'The meeting has been moved to Thursday.'
   c. State: *Tüdruku nägu läks naerule.* 'The girl started to laugh.'
   d. Addressee: *Mari rääkis Jürile kõik ära.* 'Mari told Jüri everything.'
   e. Experiencer: *Mulle meeldib siin elada.* 'I like living here.'
   f. Object of action: *Ta lootis sõpradele.* 'He counted on friends.'
   g. Object of emotions: *Mihkel on sõbrale kade.* 'Mihkel is jealous of his friend.'
   h. Without clear meaning: *Järgnege mulle.* 'Follow me.'

(4) Functions of the Estonian adessive case (Erelt, Erelt & Ross 2007: 250):
   a. Location: *Vaas on laual.* 'The vase is on the table.'
   b. Time: *Nad sõidavad neljapäeval maale.* 'They are driving to the country on Thursday.'
   c. State: *Jüri vaatas meid naerul näoga.* 'Jüri looked at us with a laughing face.'
   d. Possessor: *Maril on kaks last.* 'Mari has two children.' (Lit. 'On Mary are two children'.)
   e. Agent with finite verb forms: *See asi ununes mul kiiresti.* 'I quickly forgot about that thing.'
   f. Instrument: *Mari mängib klaveril mõnd lugu.* 'Mari is playing some tunes on the piano.'
   g. Manner: *Mari kuulas kikkis kõrvul.* 'Mari listened with her ears pricked up.'

(5) Functions of the Estonian ablative case (Erelt, Erelt & Ross 2007: 251):
   a. Source of location: *Mari võttis vaasi laualt.* 'Mari took the vase off the table.'
   b. Source: *Mari kuulis seda Jürilt.* 'Mari heard it from Jüri.'
   c. Modifier of a noun or an adjective: *Elukutselt on ta insener.* 'He is an engineer by profession.'

The postposition *peale* is an acceptable alternative only for the allative functions of direction of location, time, object of action and object of emotion. The postposition *peal* is an acceptable alternative for the adessive functions of location and instrument. The postposition *pealt* is an acceptable alternative for the ablative function of source of location. Similarly to locative cases, Estonian adpositions are also polysemous. The Dictionary of Written Estonian (Langemets et al. 2009) lists as many as 21 meanings for the postposition *peale*, 11 for both *peal* 'on' and *pealt* 'off'. A separate, polysemy account of the locative cases and the corresponding postpositional constructions is necessary, but falls outside the scope of the present study. It is hypothesised that the polysemy of both types of constructions influences the alternation between cases and postpositions and it is hoped that future research factors in polysemy when studying the alternation between near-synonymous pairs. In the present study, only the uses of the constructions that are considered as alternatives have been taken into account.

The alternating pairs can be said to be (near-)synonymous because both the locative case and the corresponding postposition render the same content; in Langacker's (1987) terminology, they profile the same relationship. Still, it is hypothesised that the variation between the synthetic case constructions and the analytic postpositional construction is not free. Even if two linguistic units do express one and the same function, they do it in different ways: they allow for a different construal of the same situation. Construal here refers to "our manifest ability to conceive and portray the same situation in alternate ways" (Langacker 2008: 43). Adhering to the tenets of the probabilistic grammar framework, it is assumed that the variation is conditioned by a set of semantic and morphosyntactic variables that will be discussed below, after the details about the data extraction and data cleaning procedures have been provided.

3. Corpus Data

3.1 Data extraction

Data for the present study were extracted from the Estonian National Corpus 2017 via Sketch Engine. The corpus contains 1.1 billion words from 3 million different documents (Kallas & Koppel 2018). ENC 2017 is an Estonian corpus of written texts that consists of the Estonian Reference Corpus (2013), Estonian Web (2013 and 2017), Estonian Wikipedia (2017) – all three sub-corpora were used for the present study. Unfortunately, there is not sufficient information available as to the size and text types of the different sub-corpora. For an approximation, the following information can be found about the sub-corpus Estonian Web 2013 (etTenTen13): uncategorised texts 35%, newspaper texts 29%, forums and blogs 23%, other types of texts 13% (Kallas, Koppel & Tuulik 2015). It is difficult to pinpoint the exact register the corpus data represents – it can vary from texts representing language use representative of spoken language (e.g. forums and blogs) to language representative of newspaper texts. For the purposes of the present study, it is crucial that the data come from a different corpus than what was used in the previous studies. Klavan 2012 and Klavan 2020 used the Morphologically Disambiguated Corpus of Estonian (MDCE[2] 2015; size 215,000 words) and the Balanced Corpus of Estonian (BCE[3] 2015; size 15 million words) for written Estonian, Klavan, Pilvik & Uiboaed 2015 used the Corpus of Estonian Dialects (CED 2015) for spoken regional dialects. The present study therefore allows to study the three alternations in a variety of Estonian that has not been used in previous studies, allowing thus to address prediction (1) of the study concerning the stability and direction of the factors across different varieties of the Estonian language.

The corpus has been automatically tagged with the Estonian Filosoft part-of-speech tagset. Automatic morphological tagging allows the extraction of case forms using the following feature query forms: [".. ad"], [".. all"], [".. abl"]. The postpositions were extracted using the

2 https://www.cl.ut.ee/korpused/morfkorpus/index.php?lang=en (last accessed: May 2021)

3 https://www.cl.ut.ee/korpused/grammatikakorpus/index.php?lang=en (last accessed: May 2021)

following query forms: [“peal”, lemma; PoS “adposition”], [“peale”, lemma; PoS “adposition”], [“pealt”, lemma; PoS “adposition”]. Using these query forms gave the results shown in Table 2[4].

Table 2. Frequency of the constructions in ENC 2017 (size: 1.1 billion words).

Construction	Result	Per million
Allative (all)	19,187,296	14,235
Adessive (ade)	30,661,120	22,748
Ablative (abl)	2,675,044	1,984
Peale	959,515	711
Peal	241,263	179
Pealt	138,049	102

Table 2 shows the global frequencies for the three exterior locative cases and the corresponding adpositions. It can be seen from the column “Per million” in Table 2 that the most frequent case in the trio is the adessive, followed by allative and ablative. For the adpositions, the most frequent one is *peale* – this reflects the fact that *peale* functions both as a preposition and a postposition in Estonian (Veismann 2006). 10,000 random lines from the total number of hits for each of the case constructions and 3,000 random lines for postpositions were generated in Sketch Engine. I saved the output as an XML file and started manually reading the lines selecting only the alternating occurrences for the final analysis. The evaluation of the alternation was based on the author’s intuition as a native speaker of Estonian. An alternation was selected only if it met the following conditions: the postposition *peale* and the allative express the functions of direction of location, time, object of action and object of emotion; the postposition *peal* and the adessive express the functions of location and instrument; the postposition *pealt* and the ablative express the function of source of location (see Section 2 above). Given that the case constructions fulfil many other functions in the language for which a postpositional construction is not a possible alternative, a large number of constructions had to be disregarded since they were not relevant for the purposes of this study. Table 3 shows how many hits I

4 The date of the query was 17 June 2018.

had to go through before reaching the desired number of 500 examples per construction. The final size of the data sample for this study is 3,000 occurrences, 500 per construction. Due to the generally much lower frequency of the postpositional variant for all three constructions, it was decided to keep the sample balanced.

Table 3. Data cleaning.

Construction	Total: cleaned	Disregarded	Sampled
Allative (all)	3,017	2,517	500
Adessive (ade)	5,148	4,648	500
Ablative (abl)	1,745	1,245	500
Peale	2,142	1,642	500
Peal	1,210	710	500
Pealt	872	372	500

3.2. Data annotation

The data were annotated according to the variables selected from previous studies, Klavan (2012, 2020) and Klavan, Pilvik & Uiboaed 2015. For the purposes of manual data annotation and the significance of these factors in previous studies, the focus is on the following: the type of construction; complexity, length (log-transformed), mobility, number and the syntactic function of the Landmark (LM) phrase; the relative position of the Trajector (TR) and Landmark phrase; the word class of the Trajector phrase; the heads of the Landmark phrase (lemmas). See Table 4 for an overview of the annotated variables and Klavan (2012: 70–92) for details and example sentences. These variables were selected for the analysis because a) they represent both morphosyntactic and semantic variables and b) they were among the strongest predictors in the previous studies on the adessive ~ *peal* alternation. In addition, these variables are used as proxies for determining the cognitive pressures related to language processing that are predicted to drive the variation between exterior locative cases and the postpositions (see prediction 3 in Section 1). More specifically, it is predicted that more complex and longer Landmark phrases are indicative of higher cognitive pressures.

Table 4. Definition of variables.

Variable name	Levels
COMPLEXITY (morphological complexity of LM)	compound, simple
CONSTRUCTION (response variable)	case, postposition
LEMMA (lemma of the LM word)	1286 lemmas
LENGTH (length of the LM phrase in syllables)	from 1 to 20 syllables long (log. transf.)
NR (number of LM)	plural, singular
MOBILITY (mobility of LM phrase)	abstract, mobile, static
POSITION (relative position btw TR and LM)	lm_tr, tr_lm
SYNFUN (syntactic function of LM)	adverbial, modifier
TRWC (word class of TR phrase)	NP, no_context, other

4. Data Analysis and Results

For the analysis of the data, mixed-effects logistic regression (Harrell 2001, Pinheiro & Bates 2000, Hosmer, Lemeshow & Sturdivant 2013) has been used in order to find out whether the choice between the case construction and the postpositional construction can be accounted for by the proposed explanatory variables in Table 4. The data were analysed using the statistical computing software *R* (version 3.6.1, R Core Team 2019). For the data analysis in this paper, the *lme4* package has been used (Bates 2014, Bates et al. 2015).

In order to arrive at the optimal model, a hypothesis-driven search for the best, i.e. the simplest yet most adequate, mixed model is used following Baayen et al. (2013). A stepwise model simplification strategy was adopted, where the minimal adequate model is selected from a set of more complex models. The stepwise progression from the maximal model (including all of the 7 variable categories as fixed effects and LM lemma as a random effect) to the minimal adequate model was made on the basis of deletion tests (F-tests or Chi-squared tests). Any redundant parameters (non-significant explanatory variables) were removed one at a time. An explanatory variable was only retained in the model if it significantly improved the fit of the model. Although there is a growing body of research that suggests against a stepwise model simplification strategy in confirmatory research designs, I decided to

take the traditional approach of seeking the most parsimonious model that accurately reflects the data in this study (Hosmer, Lemeshow & Sturdivant 2013: 89–90) and I have not reported the full models. However, model selection in regression modelling is an important methodological and theoretical issue, especially in the context of linguistic research and the decision of which approach to use for selecting variables depends on the research problem and the scientific discipline. A good recent discussion on this topic in linguistics can be found in Winter (2020: 274–280)[5].

The corpus-based models are assessed by calculating the correctly classified instances (model accuracy). Model accuracy is evaluated by two measures – percentage of overall accuracy and the C measure (Hosmer, Lemeshow & Sturdivant 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 postpositional construction is greater than or equal to 0.5 and the postpositional construction was actually observed in the data. The C measure ranges from 0.5 to 1.0 and reflects the ability of the model to discriminate between the two outcomes. The following general guidelines are given as a rule of thumb: $C = 0.5$ – no discrimination; $0.5 < C < 0.7$ – poor discrimination; $0.7 \leq C < 0.8$ – acceptable discrimination; $0.8 \leq C < 0.9$ – excellent discrimination; $C \geq 0.9$ – outstanding discrimination (Hosmer, Lemeshow & Sturdivant 2013: 177). When reporting the goodness of fit measures for the mixed-effects logistic regression models, it should be noted that the model fit has been used. This means that the models are trained and tested on the same 1,000 instances per each alternating pair.

The importance of predictors in the corpus models was assessed using decrease in the Akaike information criterion (AIC; Hosmer, Lemeshow & Sturdivant 2013: 120). 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. In a nutshell, the larger the reduction in AIC once a specific predictor is added, the more important the predictor is. One of the ways how to assess mixed-effects logistic regression models is to use

5 I am grateful to the anonymous reviewer for highlighting the issue of model selection in confirmatory linguistic research.

a Shapiro-Wilk test to assess the normality assumption of the random LEMMA-specific intercepts. Recent simulation studies have shown, however, that the misspecification of the distribution of random effects has little effect on the estimates of covariate effects (McCulloch & Neuhaus 2011, Neuhaus, McCulloch & Boylan 2013). Slight violation of the normality assumption of the random LEMMA-specific intercepts should therefore not pose problems for interpreting the main effects of the model.

4.1. The allative ~ *peale* alternation

Altogether four factors were retained (one semantic and three morphosyntactic) in the minimally adequate regression model fitted to the allative ~ *peale* data, together with the Landmark lemma as a random effect. The optimal mixed-effects logistic regression model for the allative and *peale* alternation is described by the following formula:

CONSTRUCTION ~ LOG_LENGTH + MOBILITY + SYNFUN + TRWC + (1|LEMMA)

The overall accuracy of the model is 80% and the *C* measure of 0.88 indicates that the model's discrimination between the two outcomes is excellent. The relative importance of the model predictors can be seen from Table 5, where the first column shows the order in which the three 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 5. Model comparison statistics for the mixed-effects logistic regression model for the allative ~ *peale* alternation.

	logLik	Chisq	Chi.Df	*p* value	Reduction in AIC
LEMMA	–653.19				77.9
LOG_LENGTH	–623.17	60.039	1	0.000	58.0
MOBILITY	–605.79	34.754	2	0.000	30.8
SYNFUN	–596.55	18.487	1	0.000	16.5
TRWC	–589.72	13.664	2	0.001	9.7

We can see from Table 5 that the decrease in AIC for the random effect of LEMMA is the largest (77.9), followed closely by the decrease in AIC for the fixed effect of LOG_LENGTH (58.0). The second and third fixed effect predictors, MOBILITY and SYNFUN, also make considerable contributions to the model, while the fourth fixed effect TRWC is less important, but still significant for providing a model with a better fit to the data. A slightly different ranking of predictors is obtained when instead of comparing the difference in AIC as the different predictors are added one after another to the model specification, but when models are compared with only one predictor at a time included in the model together with LEMMA as a random effect. Such a ranking still confirms that LOG_LENGTH is the strongest fixed effect predictor, but such a strategy ranks SYNFUN higher than MOBILITY followed by TRWC. The Shapiro–Wilk test indicates that the normality assumption of the random lemma-specific intercepts is violated (W = 0. 97734, *p* value < 0.001).

The coefficients in Table 6 (positive coefficient signs favour the postpositional construction and negative coefficient signs favour the case construction) indicate that the use of the allative construction increases as the length of the Landmark phrase gets longer. The allative construction is also preferred when the Landmark phrase is static and functions as a modifier in the sentence. The *peale* construction is preferred when the Landmark is a mobile entity denoted by shorter Landmark phrases that function as adverbials.

Table 6. Coefficients for the mixed-effects logistic regression model for the allative ~ *peale* alternation.

	Estimate	**Std.Error**	***z* value**	***p* value**
Intercept	1.360	0.428	3.175	0.001
LM_LENGTH	–2.305	0.344	–6.702	0.000
MOBILITY = mobile	0.484	0.209	2.304	0.021
MOBILITY = static	–1.059	0.274	–3.861	0.000
SYNFUN = modifier	–1.148	0.308	–3.725	0.000
TRWC = NP	–0.255	0.341	–0.749	0.454
TRWC = other	0.372	0.346	1.076	0.281

4.2. The adessive ~ *peal* alternation

In the minimally adequate regression model fitted to the adessive ~ *peal* data, three factors were retained (one semantic and two morpho-syntactic), together with the Landmark lemma as a random effect. Compared to the model fitted to the allative ~ *peale* alternation, TRWC is not retained in the model, but the other factors are the same. The optimal mixed-effects logistic regression model for the adessive and *peal* alternation is described by the following formula:

CONSTRUCTION ~ LOG_LENGTH + MOBILITY + SYNFUN + (1|LEMMA)

The overall accuracy of the model is 87% and the *C* measure of 0.94 indicates that the model's discrimination between the two outcomes is outstanding. The relative importance of the model predictors can be seen from Table 7, where the first column shows the order in which the three 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 7. Model comparison statistics for the mixed-effects logistic regression model for the adessive ~ *peal* alternation.

	logLik	Chisq	Chi.Df	*p* value	Reduction in AIC
LEMMA	–590.96				202.4
LOG_LENGTH	–504.60	172.729	1	0.000	170.7
MOBILITY	–465.42	78.354	2	0.000	74.4
SYNFUN	–454.75	21.339	1	0.000	19.3

We can see from Table 7 that similarly to the model fitted to the adessive ~ *peal* alternation, the decrease in AIC for the random effect of LEMMA is the largest (202.4), followed closely by the decrease in AIC for the fixed effect of LOG_LENGTH (170.7). The second fixed effect predictor MOBILITY makes also a considerable contribution to the model, while the third fixed effect SYNFUN is less important, but still significant for providing a model with a better fit to the data.

A slightly different ranking of predictors is obtained when instead of comparing the difference in AIC as the different predictors are added one after another to the model specification, but when models are compared with only one predictor at a time included in the model together with LEMMA as a random effect. Such a ranking still confirms that LOG_LENGTH is the strongest fixed effect predictor, but such a strategy ranks SYNFUN higher than MOBILITY. The Shapiro–Wilk test indicates that the normality assumption of the random lemma-specific intercepts is slightly violated (W = 0. 98971, *p* value = 0.003644).

The coefficients in Table 8 confirm the same result as for the allative ~ *peale* alternation – the use of the adessive construction increases as the length of the Landmark phrase gets longer. The adessive construction is also preferred when the Landmark phrase functions as a modifier in the sentence. The *peal* construction is preferred when the Landmark is a mobile entity denoted by shorter Landmark phrases that function as adverbials.

Table 8. Coefficients for the mixed-effects logistic regression model for the adessive ~ *peal* alternation.

	Estimate	**Std.Error**	***z* value**	***p* value**
Intercept	2.693	0.480	5.611	0.000
LM_LENGTH	–4.946	0.504	–9.801	0.000
MOBILITY = mobile	1.866	0.413	4.508	0.000
MOBILITY = static	–0.435	0.389	–1.120	0.263
SYNFUN = modifier	–1.363	0.306	–4.460	0.000

4.3. The ablative ~ *pealt* alternation

Altogether four factors were retained (one semantic and three morphosyntactic) in the minimally adequate regression model fitted to the ablative ~ *pealt* data, together with the Landmark lemma as a random effect. The optimal mixed-effects logistic regression model for the ablative and *pealt* alternation is described by the following formula:

CONSTRUCTION ~ MOBILITY + COMPLEXITY + LMTR_POSITION + LOG_LENGTH + (1|LEMMA)

The overall accuracy of the model is 86% and the *C* measure of 0.94 indicates that the model's discrimination between the two outcomes is outstanding. The relative importance of the model predictors can be seen from Table 9, where the first column shows the order in which the three 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 9. Model comparison statistics for the mixed-effects logistic regression model for the ablative ~ *pealt* alternation.

	logLik	Chisq	Chi.Df	*p* value	Reduction in AIC
LEMMA	–627.83				128.6
MOBILITY	–591.68	72.310	2	0.000	68.3
COMPLEXITY	–579.99	23.374	1	0.000	21.4
LMTR_POSITION	–577.20	5.587	1	0.018	3.6
LOG_LENGTH	–574.83	4.726	1	0.029	2.7

As with the previous two alternating pairs, we can see from Table 9 that the decrease in AIC for the random effect of LEMMA is the largest (128.6) for the ablative ~ *pealt* alternation as well. This is followed by the decrease in AIC for the fixed effect of MOBILITY (68.3), a variable that played an important role also in the other two models. The second fixed effect predictor, COMPLEXITY, did not figure in the previous models, but seems to make a considerable contribution to the model fitted to the ablative ~ *pealt* data. Another predictor that is not present in the previous models, but which makes a significant contribution for providing a model with a better fit to the data, is LMTR_POSITION. Somewhat surprisingly, LOG_LENGTH does not seem to play as decisive a role as in the other two alternations, since it only makes a very small, although a significant improvement to the model fit. I will return to the differences and similarities between the models fitted to the three alternating pairs in the discussion section of the paper.

The same ranking of predictors is obtained for the ablative ~ *pealt* alternation when instead of comparing the difference in AIC as the different predictors are added one after another to the model specification, models are compared with only one predictor at a time included in the

model together with LEMMA as a random effect. The Shapiro–Wilk test indicates that the normality assumption of the random lemma-specific intercepts is violated (W = 0. 97665, *p* value < 0.001).

The coefficients in Table 10 indicate that the ablative construction is preferred when the Landmark phrase is static, long, and complex and when the Trajector phrase precedes the Landmark phrase. The *pealt* construction is preferred when the Landmark is simple and denotes an abstract entity.

Table 10. Coefficients for the mixed-effects logistic regression model for the ablative ~ *pealt* alternation.

	Estimate	**Std. Error**	***z* value**	***p* value**
Intercept	0.782	0.431	1.813	0.069
MOBILITY = mobile	–0.463	0.264	–1.755	0.079
MOBILITY = static	–2.661	0.369	–7.199	0.000
COMPLEXITY = simple	1.139	0.299	3.808	0.000
LMTR_POSITION = tr_lm	–0.464	0.191	–2.429	0.015
LOG_LENGTH	–0.882	0.406	–2.171	0.029

4.4. The contribution of individual factors

One of the central questions for which this study seeks an answer concerns the specific factors, namely their strength and direction across the three alternating pairs. What follows, therefore, is a more detailed look at the various factors studied in the present corpus sample in order to get a better idea of how these factors contribute to the different alternations. Overall, we can see that all three models fitted to the data are deemed to provide a very good fit to the data. The prediction accuracy for all models is 80% or above and the models' discrimination between the two constructions is excellent. Even if the models are overfitting and care should be taken when using these models to make predictions about unseen data, the statistical analysis is well-fitted for describing the corpus sample and identifying the usage patterns in the data.

Based on the statistical models, the following scales can be put forward about the ranking of predictors for the three alternations:

allative ~ *peale*: LEMMA > LOG_LENGTH >
MOBILITY / SYNFUN > TRWC

adessive ~ *peal*: LEMMA > LOG_LENGTH > MOBILITY / SYNFUN

ablative ~ *pealt*: LEMMA > MOBILITY > COMPLEXITY >
POSITION > LOG_LENGTH

It is clear from the data analysis, that LEMMA, LOG_LENGTH and MOBILITY are the three variables that play a major role across all three alternating pairs, although in the ablative ~ *pealt* alternation length makes a relatively minor contribution to the model fit compared to other variables. SYNFUN is a significant predictor in the allative ~ *peale* and adessive ~ *peal* alternation, but not in the ablative ~ *pealt* alternation. LMTR_POSITION and COMPLEXITY make a significant contribution to the ablative ~ *pealt* alternation and TRWC to the allative ~ *peale* alternation. Since their variable importance as measured by reduction in AIC is of a smaller magnitude, it is concluded that they do not play a major role in the alternating pairs. Number of the LM (plural vs. single) is the only predictor that is not significant for any of the alternations.

As for the direction of the effect of the significant factors, there is considerable converging evidence both across the three alternations and in comparison with the previous studies. The data for all three alternations confirms what has been found in the previous studies that the choice between the locative case construction and the corresponding postposition depends on the length of the Landmark phrase: the longer the phrase, the more probable it is that the preferred construction is the locative case construction. In the previous studies about the adessive and *peal* alternation, the length of the Landmark phrase has been found to be one of the consistent factors to play a significant role with the longer phrases associated with the case construction (Klavan 2012, Klavan 2020, Klavan, Pilvik & Uiboaed 2015). In the present study, the same finding holds for the allative ~ *peale* alternation and the adessive ~ *peal* alternation. It can be seen from Table 11 that the mean length of the Landmark phrase is considerably longer for both the allative (5.7 syllables) and adessive (6.2 syllables) cases compared to *peale* (3.8 syllables) and *peal* (3.2 syllables) respectively. As for the ablative ~ *pealt* alternation, length of the Landmark phrase is not a significant predictor

to drive the choice between the case construction and the postpositional construction, although the general trend is the same – the length of the Landmark phrase is longer for the ablative (5.2 syllables) compared to *pealt* (4.4 syllables).

Table 11. Mean, standard deviation and the maximum of the variable LENGTH across the three alternations.

Construction	Mean of LENGTH	Standard deviation of LENGTH	Maximum of LENGTH
allative	5.7	3.6	19
peale	3.8	2.6	17
adessive	6.2	3.8	20
peal	3.2	1.9	14
ablative	5.2	3.1	17
pealt	4.4	2.7	15

Related to the variable LENGTH is the variable COMPLEXITY – compound nouns (i.e. more complex nouns) are longer than simple nouns. Contrary to the previous findings regarding the adessive ~ *peal* alternation, COMPLEXITY was not retained as a significant variable in the final models fitted to the adessive ~ *peal* dataset and the allative ~ *peale* dataset. COMPLEXITY was, however, retained in the final model fitted to the ablative ~ *pealt* dataset. Importantly, for the model fitted to the ablative ~ *pealt* dataset, multicollinearity between the variables LENGTH and COMPLEXITY is not problematic. Multicollinearity was assessed using the vif() function from the package car. Previous studies have consistently found that compound nouns clearly prefer the adessive case construction compared to the postposition *peal* construction (Klavan 2012, Klavan 2020, Klavan, Pilvik & Uiboaed 2015). The same trend can be detected for all of the three alternations if we look at the frequency counts in Table 12. The first observation to make is that in about 800 uses out of 1,000 per alternation a simple noun has been used. The second observations is that when we do have a compound noun in a sentence, at around 70% of the time the preferred construction is the case construction. Even though the variable COMPLEXITY was only retained in the final model for the ablative ~ *pealt*

alternation, it is safe to conclude based on the present findings and the previous findings that the case constructions are preferred when the Landmark noun is a compound noun.

Table 12. Frequency counts and proportions for COMPLEXITY across the three alternations.

Construction	Compound nouns	Simple nouns
allative	116 (67%)	384 (46%)
peale	56 (33%)	444 (54%)
Total	172 (100%)	828 (100%)
adessive	132 (70%)	368 (45%)
peal	55 (30%)	445 (55%)
Total	187 (100%)	813 (100%)
ablative	132 (66%)	368 (46%)
pealt	68 (34%)	432 (54%)
Total	200 (100%)	800 (100%)

Previous studies on the adessive ~ *peal* alternation have not found the variable SYNTACTIC FUNCTION to be a particularly significant factor, but the present study provides evidence that it is an important factor for the allative ~ *peale* and adessive ~ *peal* alternation. Compared to the factors LEMMA and LOG_LENGTH, the relative importance of SYNFUN remains low for the two alternating pairs, but it is retained in the final model and its effect size is comparable in scale to that of MOBILITY. As can be seen from the frequency counts in Table 13, there is a clear tendency across all three alternations for both the case construction and the postpositional construction to be used in the adverbial function (around 900 uses out of 1,000) rather than in the modifier function. However, when the locative phrase is used in the modifier function, the preferred construction is the locative case construction. The latter trend is less prominent for the ablative ~ *pealt* alternation.

Table 13. Frequency counts and proportions for SYNFUN across the three alternations.

Construction	Adverbial	Modifier
allative	432 (48%)	68 (73%)
peale	475 (52%)	25 (27%)
Total	907 (100%)	93 (100%)
adessive	410 (47%)	90 (68%)
peal	458 (53%)	42 (32%)
Total	868 (100%)	132 (100%)
ablative	437 (49%)	63 (59%)
pealt	456 (51%)	44 (41%)
Total	893 (100%)	107 (100%)

The final important piece of converging evidence pertains to the semantic factor MOBILITY. Importantly, the data for all three alternations confirms what has been found in the previous studies, namely that the choice between the locative case construction and the corresponding postposition depends on the mobility of the Landmark. For the ablative ~ *pealt* alternation, this is the factor that is ranked highest after the factor LEMMA. In the allative ~ *peale* and adessive ~ *peal* alternation it seems to play a less important role compared to the length of the Landmark phrase. For the purposes of the present study, it was decided to adopt a more intricate annotation schema for MOBILITY which resulted in a three-way division: abstract (e.g. *positsioon* 'position'), mobile (e.g. *auto* 'car') and static (e.g *tänav* 'street'). In the previous studies (Klavan 2012, Klavan, Pilvik & Uiboaed 2015), MOBILITY had only two levels. It therefore makes it more difficult to draw any direct comparisons with the previous studies where a two-fold division was used (mobile vs. static Landmarks). If we look at the frequency counts given in Table 14, we see that the adessive ~ *peal* alternation behaves differently compared to the allative ~ *peale* and ablative ~ *pealt* alternation. It behaves very similarly to what has been put forward in the previous studies – in the majority of the uses (in roughly 900 occurrences out of 1,000) either a mobile or a static Landmark has been used, only 121 uses have an abstract Landmark. For the allative ~ *peale* and ablative ~ *pealt* alternation, as many as 400 uses out of 1,000 have

an abstract Landmark. However, when the Landmark is abstract, in the allative ~ *peale* alternation the preferred construction is the allative construction, but in the ablative ~ *pealt* alternation, the preferred construction is the *pealt* construction. For static and mobile Landmarks, the trend is the same across all three alternations – the postpositonal constructions are preferred with mobile Landmarks and locative cases with static Landmarks.

Table 14. Frequency counts and proportions for MOBILITY across the three alternations.

Construction	Abstract	Mobile	Static
allative	231 (57%)	136 (33%)	133 (70%)
peale	172 (43%)	271 (67%)	57 (30%)
Total	403 (100%)	407 (100%)	190 (100%)
adessive	84 (70%)	86 (22%)	330 (68%)
peal	37 (30%)	309 (78%)	154 (32%)
Total	121 (100%)	395 (100%)	484 (100%)
ablative	187 (43%)	142 (41%)	171 (77%)
pealt	244 (57%)	204 (59%)	52 (23%)
Total	431 (100%)	346 (100%)	223 (100%)

4.5. The contribution of individual words

A consistent and significant finding across the three alternations in the present study is that individual words (represented in the analysis by the random effect factor LEMMA) make a very prominent contribution towards accounting for the variation found in the corpus data. It can be argued, based on the reduction in Akaike information criterion for the three models presented in Sections 4.1–4.3 above, that individual words are considerably more important for the adessive ~ *peal* alternation (reduction in AIC: 202.4) than for the ablative ~ *pealt* alternation (reduction in AIC: 128.6) and the allative ~ *peale* alternation (reduction in AIC: 77.9). This result ties in with the number of different words that occur in the 1,000-sentence corpus sample for each alternation. For the adessive ~ *peal* alternation, the number of different lemmas is 438 compared to 544 for ablative ~ *pealt* and 611 for allative ~ *peale*. These

numbers indicate that the variation of words is somewhat more limited for the adessive ~ *peal* alternation and it is clear why the random effect for LEMMA is bigger in the model fitted to this alternation.

Table 15 lists the 72 word lemmas that appear in all three datasets together with the random-effect values for these words according to the three models reported in Sections 4.1–4.3. The positive values represent a word bias towards the postpositional construction and negative values a bias towards the case construction. Technically the numerical values for the random effect of lemmas shown in Table 15 are not estimated parameters for the statistical model, but 'best linear unbiased predictors (BLUPS)' (Pinheiro & Bates 2000:71).

Table 15. Random effect values of the corpus models.

LEMMA	allative ~ peale	adessive ~ peal	ablative ~ pealt	LEMMA	allative ~ peale	adessive ~ peal	ablative ~ pealt
aken	–0.350	–0.889	0.800	meeter	0.567	1.794	1.093
ala	–0.329	–1.597	–0.803	meri	–0.600	0.033	–0.533
alus	–0.117	0.363	–0.304	mis	–0.262	–1.255	–0.210
ametikoht	–0.570	–0.550	–0.532	nägu	–0.621	–1.651	–1.091
areen	–0.226	–0.146	–0.532	paber	–0.805	–0.019	–0.302
aste	–0.056	–0.607	1.135	pann	–0.297	–1.304	–0.383
auto	0.915	0.232	0.553	piir	0.903	0.427	–0.297
buss	0.652	0.076	–1.543	pilt	–0.064	–0.774	–0.175
eriala	0.104	0.087	–0.882	pind	–0.307	–0.566	0.317
ise	–0.143	0.202	–1.005	pink	–0.510	–1.244	–0.723
jalg	0.103	–0.984	–1.689	plaat	–0.533	–0.797	–0.436
kaart	–0.510	0.490	–0.131	plats	0.068	1.040	–0.570
kallas	–0.248	0.081	0.174	positsioon	–0.193	–0.704	–1.797
kanal	1.151	–0.252	1.884	põld	–0.079	0.654	–0.626
kapp	0.779	0.232	0.384	põlv	–0.262	0.518	0.384
keha	–0.602	–0.984	–1.147	põrand	–0.593	0.049	–0.777
kivi	–0.262	0.405	0.506	rada	0.473	–0.387	–0.572
koduleht	0.052	0.034	–2.620	ratas	–0.484	–1.428	–0.500
koht	1.077	–1.320	1.626	rind	0.213	–1.141	–0.980
kolv	0.213	0.159	0.452	rong	0.867	0.256	–0.383
korrus	–0.457	–0.297	–0.474	samm	0.518	–0.711	–0.617
kõht	0.486	0.470	0.762	see	0.665	0.122	0.311

LEMMA	allative ~ peale	adessive ~ peal	ablative ~ pealt	LEMMA	allative ~ peale	adessive ~ peal	ablative ~ pealt
käsi	0.590	0.138	0.423	sein	0.060	0.896	–0.860
külg	–0.723	–0.984	0.394	sõit	0.023	–0.429	0.852
küünal	0.213	–1.165	0.506	süda	–0.069	0.159	–1.177
laev	–0.428	0.634	–1.650	tasand	–0.183	–1.003	–0.389
laht	–0.133	–0.711	–0.267	tase	–1.233	–0.554	–0.975
laud	–0.298	–0.183	–0.204	tee	1.099	–1.622	–0.740
lava	–1.019	–1.306	1.121	toode	–0.186	–1.010	0.720
lehekülg	–0.233	0.220	–2.035	tuli	–0.248	–0.712	–0.694
leht	–0.861	–1.366	–2.665	turg	–0.752	–2.131	–1.410
lennuk	0.403	0.350	–0.588	tänav	–0.651	–1.590	–1.352
lett	–0.048	1.362	–0.453	uks	–0.125	–1.136	–0.207
liin	–0.248	–0.524	–0.212	veebileht	–0.173	–0.271	–0.943
liiv	0.795	0.427	–0.238	voodi	–0.570	0.393	–0.383
maantee	0.543	–0.472	–0.339	väljak	–0.474	–1.340	–0.471
maastik	–0.083	–0.849	–0.392	õlg	–1.131	0.076	0.506
mees	–0.092	0.159	0.506	õu	0.806	1.863	–0.321

The random-effect values in Table 15 demonstrate that there are some words that behave the same way across all three alternations – they have a positive or negative value in all three models. For example, words that have a negative value and are therefore biased towards the case construction are *ala* ‘field’, *leht* ‘leaf; page’, *mis* ‘what’, *nägu* ‘face’, *pann* ‘pan’, *pilt* ‘photo’, *pink* ‘bench’, *positsioon* ‘position’, *ratas* ‘wheel’, *turg* ‘market’, *tänav* ‘street’, *uks* ‘door’, *väljak* ‘square’. Words that have a positive value and are biased towards the postpositional construction irrespective of the alternating pair include *auto* ‘car’, *kanal* ‘channel’, *kapp* ‘wardrobe’, *kõht* ‘stomach’, *käsi* ‘hand’, *meeter* ‘metre’. At the same time, there are also words that have a very different bias depending on the construction. For example, *buss* ‘bus’, *koduleht* ‘web-page’ and *lehekülg* ‘page’ have a comparatively high negative value in the ablative ~ *pealt* alternation (the preferred construction being the ablative case), but a positive value in the other two alternations. Hence, the tendency is to find in the data *bussilt* rather than *bussi pealt* (‘off the bus’), but we prefer *bussi peale* rather than *bussile* (‘onto the bus’). The words *koht* ‘place’ and *tee* ‘road’ have a strong negative value for the adessive ~ *peal* alternation (bias towards *kohal* and *teel*), but a strong

positive value for the allative ~ *peale* alternation (bias towards *koha peale* and *tee peale*). It is clear that specific words are biased towards one or the other construction to varying degrees. There are both crucial similarities and interesting differences when we look at the biases of the words across the three alternations.

5. Discussion

The contribution of the present paper to the study of external locative case constructions and the corresponding postpositional constructions in present-day Estonian are twofold. First of all, the present study takes stock on all three alternations simultaneously – previous studies have only focused on the adessive ~ *peal* alternation. Second, the present study looks at the alternations in a very large dataset based on web-based texts – a register that has not been studied previously in the context of these alternations. As such, the present study adds to the body of knowledge how probabilistic choice making processes for Estonian morpho-syntactic alternations differ across varieties of the same language.

One of the most important conclusions to be taken away from this study is that the adessive ~ *peal* alternation and the allative ~ *peale* alternation exhibit similar usage patterns compared to each other and compared to the previous findings on the adessive ~ *peal* alternation in other registers; the ablative ~ *pealt* alternation, however, shows usage patterns that differ from the other two alternations. It may be hypothesised that these differences are likely related to the different status of the alternations in Estonian grammar. In the alternating pairs allative ~ *peale* and adessive ~ *peal* the case constructions carry many other meanings in addition to the locative meaning for which the postpositional constructions are viable alternatives. The present study does not discuss the polysemy of the alternating constructions, but the results of the present study indicate that future work on all three constructions should take into account polysemy when looking at the synonymy between the alternations.

Consistently with the previous findings about the adessive ~ *peal* alternation (Klavan 2012, Klavan 2020, Klavan, Pilvik & Uiboaed 2015), the present study confirms that MOBILITY, LENGTH and COMPLEXITY of the Landmark phrase play a significant role for the

Estonian morphosyntactic alternations between external locative cases and the corresponding postpositions, although the strength and direction of the variables differs across the alternations. The present study adds to the body of evidence found for other Finno-Ugric languages on a different set of locative cases – the interior locative cases and the corresponding adpositions in the Saami (Bartens 1978) and Finnish language (Ojutkangas 2008). In line with Bartens (1978), the present study shows that the postpositional constructions are used together with smaller, manipulable things. In fact, the variable MOBILITY was one of the variables retained in all three models fitted to the three alternations. For static and mobile Landmarks, the trend is the same across all three alternations – the postpositonal constructions are preferred with mobile Landmarks and locative cases with static Landmarks. However, when the Landmark is abstract, the preferred construction in the allative ~ *peale* alternation is the allative construction, but in the ablative ~ *pealt* alternation, the *pealt* construction.

Probabilistic grammars are surprisingly stable in a cross-variety perspective. The present study confirms the findings of the previous studies that MOBILITY and LENGTH in addition to the random effect variable LEMMA are three very important factors in the alternations between exterior locative cases and postpositions. As with previous studies on syntactic alternations from a probabilistic grammar perspective (e.g. Grafmiller et al. 2018), we do not see any reversals in effect directions. For example, the constraints of MOBILITY and LENGTH have the same qualitative effect across varieties – present-day written Estonian (Klavan 2012, 2020), nonstandard spoken Estonian (Klavan, Pilvik & Uiboaed 2015) and web texts in Estonian (the present study). In addition, there are no reversals in effect directions for these factors across the three different alternations. What we do see, however, are interesting quantitative differences with regard to the effect size of the constraints on variation depending on the variety and the specific alternation.

The typological contribution of the present study is that it looks at a series of morphosyntactic alternations that do not concern changes in the word order like the dative alternation (*John gave the book to Mary* vs. *John gave Mary the book*) and the genitive alternation (*John's mother* vs. *the mother of John*) in English. One of the constraints that is exhibited fairly constantly across varieties and alternations in English

is constituent length (Grafmiller et al. 2018). True, length of the Landmark phrase is an important factor for the Estonian morphosyntactic alternations, but the reasoning behind it is different. For the English language we can see how constituent length is connected with the principle of end-weight (Wasow & Arnold 2003) – language users tend to place longer, heavier constituents after shorter ones. For the alternations between exterior cases and postpositions, length is connected with the principle of economy (Haiman 1983) and Zipf's "principle of least effort" (Zipf 1935). If we take the postpositional construction to be more complex than the locative case construction, it may be argued that language users avoid making an already long Landmark phrase even longer and opt for the shorter case inflection instead.

Length of the Landmark phrase is definitely an important factor, but even the present study demonstrates that even though there are no reversals in the effect direction – the constant finding being that the longer the Landmark phrase, the more probable the case construction – there are differences in the effects size across the alternations and across different studies. In the present study, length contributes much less to the final model in the ablative ~ *pealt* alternation compared to the allative ~ *peale* and adessive ~ *peal* alternation. Klavan (2020) reports the results of a mixed-effects logistic regression model fitted to a corpus sample that consists of 900 occurrences of the adessive ~ *peal* alternation in present-day written Estonian. Importantly, according to this mixed-effects model length of the Landmark phrase contributed more to the model fit than the random effect variable LEMMA. At the same time, Klavan (2012) reports the findings of an acceptability judgement task where the length of the Landmark phrase was intentionally manipulated together with the type of the Landmark. According to the experimental results, length was not a statistically significant factor. It should be pointed out, however, that due to the "laboratory" setting of the experiment, length was transformed into a binary variable with short and long phrases, taking away the broader range of variation exhibited in naturally occurring corpus data.

As for future work, an interesting line of investigation is the observation made by Grafmiller et al. (2018) who note that different alternations differ as to how amenable they are to probabilistic effects, i.e. different alternations are either more or less variable in a cross-variety perspective. The present study has provided evidence that the

adessive ~ *peal* alternation exhibits similar probabilistic patterns as found in previous studies on this alternation in other varieties. Currently, there are no studies that look at the other two alternations in a cross-variety perspective. The next step for this line of study is to include both written and spoken data for all three alternations to further explore the prediction about the stability and direction of the various factors across different varieties of Estonian. Only by exploring systematically morphosyntactic alternations in a morphologically rich language such as Estonian can we hope to advance the theoretical and empirical discussion of probabilistic grammar analysis and the nature and limits of grammatical variation.

6. Conclusion

The present study demonstrates that there are both similarities and differences in the morphosyntactic knowledge on the part of Estonian speakers as pertains to the three alternations between exterior locative cases (allative, adessive, ablative) and the corresponding postpositions (*peale*, *peal*, *pealt*). By exploring a large, manually annotated dataset of Estonian web texts (3,000 occurrences in total), it has been possible to determine the probabilistic variation patterns. The data were manually annotated for the following variables: complexity, length, number, mobility and the syntactic function of the Landmark phrase; the relative position of the Landmark and Trajector phrase; the word class of the Trajector phrase; and individual word lemmas. Mixed-effects logistic regression was used on the annotated corpus sample to capture the speakers' multivariate and probabilistic knowledge quantitatively.

The accuracy of the mixed-effects logistic regression models fitted to the data vary from 80% ($C = 0.88$) for the allative ~ *peale* alternation to 86% ($C = 0.94$) for the ablative ~ *pealt* alternation and 87% ($C = 0.94$) for the adessive ~ *peal* alternation. The models provide a very good fit and confirm the relevance of three factors across the three alternation: length and mobility of the Landmark phrase and the individual lemmas. Longer phrases are predictive for the locative case construction, as are static entities that denote places (e.g. *turg* 'market', *tänav* 'street', *väljak* 'square'). The study also shows that considerable variation in the data can be explained with the inclusion of individual lemmas. Individual

words are biased towards the case construction or the postpositional construction to various degrees and this bias can be either the same for all three alternations, or words can have a different bias depending on the specific alternation.

The present study tested three specific predictions put forward in the context of probabilistic grammar (cf. Grafmiller et al. 2018). First, by replicating the previous studies carried out on present-day standard written Estonian (Klavan 2012, 2020) and non-standard spoken Estonian (Klavan, Pilvik & Uiboaed 2015), it was shown that the influence of length and mobility of the Landmark phrase on the morphosyntactic variation between exterior locative cases and the corresponding postpositions is relatively stable in terms of the direction of those factors in different varieties of Estonian. Second, the strength of different factors on the speakers' choices varies by the type and frequency of the constructions – the alternation between adessive ~ *peal* is more frequent compared to the other two alternations and the results of the present study show that it was affected by three variables compared to the four variables that played a role in the other two alternations. Furthermore, the alternation between ablative ~ *pealt*, which is prominently less frequent compared to the other two alternations, is affected by a slightly different combination of variables compared to allative ~ *peale* and adessive ~ *peal*. Third, the variation in the use of exterior locative cases and the postpositions is driven by stylistic preferences among registers and speakers, situational forces (e.g. the dialectal differences demonstrated in Klavan, Pilvik & Uiboaed 2015) and by cognitive pressures related to language processing (e.g. exterior locative cases are preferred with longer and more complex Landmark phrases). Overall, the multivariate analysis of corpus data shows that the grammatical knowledge of Estonian exterior cases and the corresponding postpositions is probabilistic and regulated by both morphosyntactic and semantic factors.

Acknowledgements

The work of the author was supported by a grant from the Estonian Research Council (PUT1358 "The Making and Breaking of Models: Experimentally Validating Classification Models in Linguistics"). The author would like to thank the two anonymous reviewers for their useful comments and suggestions.

References

Abdulrahim, Dana. 2013. *A corpus study of basic motion events in Modern Standard Arabic.* Edmonton: University of Alberta dissertation. http://hdl.handle.net/10402/era.33921. (20 February, 2015.)

Anderson, John M. 2006. *Modern grammars of case*. Oxford: Oxford University Press.

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

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

Bartens, Raija. 1978. *Synteettiset ja analyyttiset rakenteet lapin paikanilmauksissa* (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 May, 2015.)

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

Bod, Rens, Jennifer Hay & Stefanie Jannedy (eds.). 2003a. *Probabilistic linguistics*. Cambridge, MA: MIT Press.

Bod, Rens, Jennifer Hay & Stefanie Jannedy (eds.). 2003b. Introduction. In Rens Bod, Jennifer Hay & Stefanie Jannedy (eds.), *Probabilistic linguistics*, 1–10. Cambridge, MA: MIT Press.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina & R. Harald Baayen. 2007. Predicting the dative alternation. In Gerlof Bouma, Irene Krämer & Joost Zwarts (eds.), *Cognitive foundations of interpretation*, 69–94. Amsterdam: Royal Netherlands Academy of Science.

Bresnan, Joan & Jennifer Hay. 2008. Gradient grammar: An effect of animacy on the syntax of give in New Zealand and American English. *Lingua* 118(2). 245–259. https://doi.org/10.1016/j.lingua.2007.02.007.

Bresnan, Joan & Marilyn Ford. 2010. Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language* 86(1). 168–213.

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*, 75–96. Berlin: Mouton de Gruyter.

Bybee, Joan. 2006. From usage to grammar: The mind's response to repetition. *Language* 82(4). 711–733. https://doi.org/10.1353/lan.2006.0186.

Bybee, Joan & Paul Hopper. 2001. *Frequency and the emergence of linguistic structure*. Amsterdam: Benjamins. https://doi.org/10.1075/tsl.45.

Divjak, Dagmar. 2010. *Structuring the lexicon: A clustered model for near-synonymy* (Cognitive Linguistics Research 43). Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110220599.

Divjak, Dagmar & Antti Arppe. 2013. Extracting prototypes from exemplars: What can corpus data tell us about concept representation? *Cognitive Linguistics* 24(2). 221–274. https://doi.org/10.1515/cog-2013-0008.

Divjak, Dagmar, Antti Arppe & Ewa Dąbrowska. 2016. Machine meets man: Evaluating the psychological reality of corpus-based probabilistic models. *Cognitive Linguistics* 27(1). 1–33. https://doi.org/10.1515/cog-2015-0101.

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

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

Grafmiller, Jason, Benedikt Szmrecsanyi, Melanie Röthlisberger & Benedikt Heller. 2018. General introduction: A comparative perspective on probabilistic variation in grammar. *Glossa: A Journal of General Linguistics* 3(1). 1–10. http://doi.org/10.5334/gjgl.690.

Gries, Stefan Th. 2003. *Multifactorial analysis in corpus linguistics: A study of particle placement*. New York: Continuum Press.

Haiman, John. 1983. Iconic and economic motivation. *Language* 59(4). 781–819.

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

Heller, Benedikt, Benedikt Szmrecsanyi & Jason Grafmiller. 2017. Stability and fluidity in syntactic variation world-wide: the genitive alternation across varieties of English. *Journal of English Linguistics* 45(1). 3–27. https://doi.org/10.1177/0075424216685405.

Heller, Benedikt & Benedikt Szmrecsanyi. 2019. Possessives world-wide: Genitive variation in varieties of English. In Nuria Yáñez-Bouza, Emma Moore, Linda van Bergen & Willem B. Hollmann (eds.), *Categories, constructions, and change in English syntax*, 315–335. Cambridge: Cambridge University Press.

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

Kallas, Jelena, Kristina Koppel & Maria Tuulik. 2015. Korpusleksikograafia uued võimalused eesti keele kollokatsioonisõnastiku näitel. *Eesti Rakenduslingvistika Ühingu aastaraamat* 11. 75–94. http://dx.doi.org/10.5128/ERYa11.05.

Kallas, Jelena & Kristina Koppel. 2018. *Eesti keele ühendkorpus 2017*. Center of Estonian Language Resources. https://doi.org/10.15155/3-00-0000-0000-0000-071E7L.

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. 2020. Pitting corpus-based classification models against each other: a case study for predicting constructional choice in written Estonian. *Corpus Linguistics and Linguistic Theory* 16(2). 363–391. https://doi.org/10.1515/cllt-2016-0010.

Klavan, Jane & Dagmar Divjak. 2016. The cognitive plausibility of statistical classification models: Comparing textual and behavioral evidence. *Folia Linguistica* 50(2). 355–384. https://doi.org/10.1515/flin-2016-0014.

Klavan, Jane, Maarja-Liisa Pilvik & 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.

Klavan, Jane & Ann Veismann. 2017. 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'. *Eesti ja soome-ugri keeleteaduse ajakiri. Journal of Estonian and Finno-Ugric Linguistics* 8(2). 59–91. https://doi.org/10.12697/jeful.2017.8.2.03.

Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia, PA: University of Philadelphia Press.

Labov, William. 1982. Building on empirical foundations. In Winfred Lehmann & Yakov Malkie (eds.), *Perspectives on historical linguistics*, 17–92. Amsterdam, Philadelphia: Benjamins. https://doi.org/10.1075/cilt.24.06lab.

Langacker, Ronald W. 1987. *Foundations of cognitive grammar. Volume I: Theoretical prerequisites*. Stanford: Stanford University Press.

Langacker, Ronald W. 2008. *Cognitive grammar. A basic introduction.* Oxford: Oxford University Press.

Langemets, Margit, Mai Tiits, Tiia Valdre, Leidi Veskis, Ülle Viks & Piret Voll. 2009. *Eesti kirjakeele seletussõnaraamat. 4 P–R.* Tallinn: Eesti Keele Sihtasutus.

Lyons, John. 1977. *Semantics*, vol. 2. Cambridge: Cambridge University Press.

Matsumura, Kazuto. 1994. Is the Estonian adessive really a local case? *Journal of Asian and African Studies* 46/47. 223–235.

McCulloch, Charles E. & 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. https://doi.org/10.1214/11-STS361.

Neuhaus, John M., Charles E. McCulloch & 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. https://doi.org/10.1002/sim.5682.

Ojutkangas, Krista. 2008. Mihin suomessa tarvitaan sisä-grammeja? *Virittäjä* 112(3). 382–400.

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

R Core Team. 2019. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Röthlisberger, Melanie, Jason Grafmiller & Benedikt Szmrecsanyi. 2017. Cognitive indigenization effects in the English dative alternation. *Cognitive Linguistics* 28(4). 673–710. https://doi.org/10.1515/cog-2016-0051.

Szmrecsanyi, Benedikt. 2013. The great regression: genitive variability in Late Modern English news texts. In Kersti Börjars, David Denison & Alan Scot (eds.), *Morpho-syntactic categories and the expression of possession*, 59–88. Amsterdam: Benjamins. https://doi.org/10.1075/la.199.03szm.

Szmrecsanyi, Benedikt, Jason Grafmiller, Benedikt Heller & Melanie Röthlisberger. 2016. Around the world in three alternations: Modeling syntactic variation in varieties of English. *English World-Wide* 37(2). 109–137. https://doi.org/10.1075/eww.37.2.01szm.

Tagliamonte, Sali A. 2011. *Variationist sociolinguistics: Change, observation, interpretation*. Oxford/New York: Wiley-Blackwell.

Veismann, Ann. 2006. *Peale* ja *pärast*. *Emakeele Seltsi Aastaraamat* 51. 170–183.

Veismann, Ann & Mati Erelt. 2017. Kaassõnafraas. In Mati Erelt & Helle Metslang (eds.), *Eesti keele süntaks*, 446–462. Tartu: Tartu Ülikooli kirjastus.

Wasow, Thomas & Jennifer Arnold. 2003. Post-verbal constituent ordering in English. In Günter Rohdenburgand & Britta Mondorf (eds.), *Determinants of grammatical variation in English*, 119–154. Berlin: Mouton de Gruyter.

Winter, Bodo. 2020. *Statistics for linguists: An introduction using R*. London: Routledge. https://doi.org/10.4324/9781315165547.

Zipf, George K. 1935. *The psycho-biology of language: An introduction to dynamic philology*. Boston: Houghton Mifflin.

Kokkuvõte. Jane Klavan: Eesti keele väliskohakäänete ja kaassõnade *peal, peale, pealt* kasutus eestikeelses veebis. Tõenäosusliku grammatika raamistikus eeldatakse, et grammatiline teadmine hõlmab endas tõenäosuslikku komponenti ja et see tõenäosuslik komponent pärineb suures osas keele kasutuse kogemusest. Sellistelt põhimõtetelt lähtuvate uurimuste eesmärgiks on mõõta grammatilise teadmise ulatust ja olemust nagu see peegeldub keelelises varieeruvuses. Esitan suuremahulise korpusuurimuse eesti keele väliskohakäänete ja nendega rööpselt tarvitatavate kaassõnade (*peale*, *peal*, *pealt*) paralleelsest kasutusest eestikeelsetel veebilehtedel. Korpusandmete multifaktoriaalne analüüs näitab, et grammatiline teadmine sellest rööpsest kasutusest on tõenäosuslik ja et seda reguleerivad nii morfosüntaktilised kui semantilised tegurid.

Märksõnad: kohakäänded, kaassõnad, süntaktiline varieerumine, keele varieerumine, tõenäosuslik grammatika, segamudelid, eesti keel