Jane Klavan* and Ole Schützler*

The complexity principle and the morphosyntactic alternation between case affixes and postpositions in Estonian

https://doi.org/10.1515/cog-2021-0114
Received October 18, 2021; accepted June 19, 2023; published online July 7, 2023

Abstract: This paper investigates three morphosyntactic alternations in Estonian – those between the exterior locative cases allative, adessive and ablative and the corresponding postpositions *peale* 'onto', *peal* 'on' and *pealt* 'off'. Based on the Complexity Principle (e.g., Rohdenburg, Günter. 2002. Processing complexity and the variable use of prepositions in English. In Hubert Cuyckens & Günter Radden (eds.), *Perspectives on prepositions*, 79–100. Tübingen: Niemeyer), we expect cognitively more complex constructions to use more explicit (i.e., morphologically more substantial) marking by means of a postposition. Further, we expect variation to be conditioned similarly in all three semantic categories. For each of the three alternations, a random sample of the two outcomes (case vs. postposition) from the Estonian National Corpus is used, resulting in 3,000 data points. Using properties of the Landmark phrase as independent variables in Bayesian mixed-effects logistic regression models, we predict the choice of postpositions over case-marked realisations. Of the patterns found, only the frequency-related one supports our complexity-related hypothesis. We conclude that the Complexity Principle, in its general form, has little explanatory power for the Estonian constructions at hand and, in particular, that the derived principle of 'analytic support' is not generally applicable. We show, however, that the grammatical knowledge of Estonian exterior locative cases and the corresponding postpositions is regulated by our three factors in a relatively uniform way.

Keywords: alternations; Bayesian regression; complexity principle; Estonian

1 Introduction

Studying the alternation between case affixes and postpositions in three different but semantically related locative constructions allows us to look at a component of

***Corresponding authors: Jane Klavan**, Department of English Studies, University of Tartu, 51007, Ülikooli 18, Tartu, Estonia; and **Ole Schützler**, Institute of British Studies, Leipzig University, Leipzig, Germany, E-mail: jane.klavan@ut.ee (J. Klavan), ole.schuetzler@uni-leipzig.de (O. Schützler)

human languages that combines three essential aspects – syntax, semantics and morphology. Our focus on morphosyntax and morphosemantics sets our study apart from previous studies on alternations that have mainly focused on syntax and word order in a single semantic domain, e.g., the genitive and dative alternations in English (e.g., Antila et al. 2010; Bresnan 2007; Bresnan et al. 2007; Bresnan and Ford 2010; Grafmiller et al. 2018; Heller et al. 2017; Rosenbach 2014; Wolk et al. 2013). By testing the fit between the variation in the Estonian data and the assumptions made by the Complexity Principle (or CP, for short; e.g., Rohdenburg 2002), we explore whether or not those assumptions apply to languages (like Estonian) that are typologically remote from English – in effect, we thus test the validity of the CP as a general, cross-linguistic principle. The CP states that more explicit grammatical alternatives tend to be preferred in cognitively more complex environments (Rohdenburg 1996).

In our paper, we focus on three morphosyntactic alternations that involve the choice between a postposition and a case affix for marking a locative function. We analyse these phenomena against the background of the CP because, like Rohdenburg (1996), we take the stance that grammatical explicitness is not only about ideal word order and the avoidance of syntactic discontinuities (cf. Hawkins 1990, 1992, 1994). In our case, both variants (case affix; postposition) are placed at the end of the governing term, so the postpositional variant is taken to be more explicit not on positional grounds but because it is phonologically more bulky and prominent than the equivalent case affix (hence a more effective signal); it is also of a higher morphological rank, being a free element rather than part of an inflected word form; and, semantically, it is more specific in meaning than the far more general and polysemous corresponding case affix. Following the Complexity Principle, we predict that the postpositional variant will be preferred if the construction as a whole is classified as more complex. What it means for a construction to be 'cognitively complex', and what it means to be 'grammatically explicit', will be explained in more detail in Section 2.2.

The three exterior locative constructions in our study are labelled *lative*, *locative* and *separative*. Lative and separative are non-static spatial functions that mean, respectively, 'move towards a position in contiguity with it' and 'move from a place'. Locative is a static spatial function that broadly means 'to be contiguous with a place'. In discussing the participants involved in such spatial functions, 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 relevant; 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, corresponding to the entity inflected for the

genitive case if followed by the postposition *peale, peal* and *pealt*. The concepts of Trajector and Landmark are closely related to Talmy's (1983, 2000) notions of Figure and Ground. The concepts of Figure-Ground and Trajector-Landmark primarily pertain to the same phenomena. Although the terminology of Figure-Ground is employed more frequently in spatial language studies, the present study uses Langacker's (1987, 2008) terminology. We agree with Langacker (1987: 231–237), who presents arguments for treating the Trajector-Landmark asymmetry as a particular case of Figure-Ground alignment. The Figure-Ground alignment refers to a more general mechanism in human cognition, while Trajector-Landmark applies more specifically to linguistic expressions. The point of departure in the present study is the linguistic expression.

We treat the three semantic relations of lative, locative and separative as more specific manifestations of a more general exterior locative concept. Examples 1–3 highlight how the spatial functions of lative, locative and separative can be expressed with a postposition and a case affix. The first option for expressing each of the three propositions is to use the case affix, which results in the word forms *lauale, laual, laualt*. The second option is to use the adpositions *peale* 'onto', *peal* 'on', *pealt* 'off' together with the noun, e.g., *laua peale, laua peal, laua pealt*.

(1) LATIVE

Paneb	raamatu	{lauale	/	laua	peale. }
Put-PRS.3SG	book.SG.GEN	table.SG.ALL	/	table.SG.GEN	onto

"He/She puts the book on(to) the table."

(2) LOCATIVE

Raamat	on	{laual	/	laua	peal. }
book.SG.NOM	be-PRS.3SG	table.SG.ADE	/	table.SG.GEN	on

"The book is on the table."

(3) SEPARATIVE

Võtab	raamatu	{laualt	/	laua	pealt. }
take-PRS.3SG	book.SG.GEN	table.SG.ABL	/	table.SG.GEN	from on

"He/She takes the book from the table."

Although many languages have both adpositions and case affixes (Hagège 2010: 9), only a few languages have a system to mark the same function with an adposition and case affix in a parallel manner. Estonian is one such typologically interesting language. Even though case affixes and adpositions serve the same purpose of function marking by setting up a link between the term they govern and the head (Hagège 2010: 37), they show many phonological, morphological, syntactic, diachronic, semantic and pragmatic differences. The former are more frequent in a

language that uses both case affixes and adpositions. A common explanation is that case affixes are shorter (perhaps a reflection of the general principle of length being inversely proportional to frequency) and that they are the preferred markers of core functions that occur very frequently in a language (Hagège 2010: 29). This also holds for the three case affixes studied in the present paper – both the allative and adessive are markers of core functions (see Appendix A for examples). One of the morphological differences between case affixes and adpositions is case doubling (Hagège 2010: 26). In Estonian, as in other inflectional languages, case morphemes are suffixed to other components of the governed term (e.g., its adjectival attributes and determiners) in addition to the head noun, while adpositions occur only once. It is commonly proposed that languages use case affixes to express more abstract relations and adpositions to express more concrete ones (Comrie 1986; Erelt et al. 1995: 33–34; Hagège 2010: 38; Lestrade 2010).

Our study sets out to explore the homogeneity (or heterogeneity) of these patterns for three related constructions in Estonian to establish, firstly, if they support the notion of a superordinate, more general locative 'super-construction' – the three more specific constructions would then alternate in similar ways. Secondly, we test the patterns against concrete hypotheses based on the Complexity Principle. Putting both aspects together, we hope to contribute both to theory building for a specific aspect of Estonian grammar and to a better understanding (and reassessment) of a theoretical approach that has thus far been mainly applied to English, particularly with a view to its broader applicability in typologically different languages.

2 Theoretical background and hypotheses

2.1 The alternation between exterior locative constructions in Estonian

Estonian has 14 nominal cases, both in singular and plural; three are called "exterior locative cases": allative, adessive, and ablative. Estonian reference grammars usually make very general claims on the lines that the meaning of adpositions is more concrete and specific, while the meaning of locative cases is more abstract and their range of uses broader (Erelt et al. 1995: 33–34, 2007: 191; Veismann and Erelt 2017: 446). The only pair of alternations that has been studied quantitatively using state-of-the-art statistical analysis is the adessive ~ *peal* alternation (Klavan 2012, 2020; Klavan et al. 2015; Klavan and Veismann 2017).

The three Estonian (exterior) locative case affixes fulfil many relatively abstract functions besides expressing spatial relations. For instance, it is more frequent for

the case affix to express temporal relations or addressees, experiencers, possessors, agents, and sources than location (see Appendix A for examples; Klavan et al. 2020). 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 only for the adessive functions of location and instrument; and the postposition *pealt* is an acceptable alternative only for the ablative function of source of location. The domain of variation is the locative function, but the functional versatility of case affixes has implications for our quantitative analysis (as explained in Section 3.1).

The estonian postpositions *peale, peal, pealt* are complex adpositions: "a word-type in which one of the constituent parts is an adposition, and the other is a case affix, in a form required by the adposition" (Hagège 2010: 38). This particular make-up yields further support to our working assumption that the three relations (and their realisations) can perhaps be treated as constructions within a more general locative super-construction. The postpositions *peale, peal,* and *pealt* are linked to the governed term by a genitive case. Overall, complex adpositions are quite pervasive in Finnic languages. It is vital to remember that an adpositional phrase in Estonian typically does not consist of only an adposition and a noun but includes case affixes on the noun and very often on the adposition itself, e.g. the allative case affix on *pea-le*, the adessive case affix on *pea-l* and the ablative case affix on *pea-lt*.

Syntactically, the Estonian locative cases and postpositions typically take the role of an adverbial (as in *laual/laua peal* 'on the table' in example 4) or adverbial modifier (as *vaas laual/vaas laua peal* 'the vase on the table' in example 5) (Erelt et al. 1995: 58).

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

(5) Vaas {laual / laua peal} on ilus.
vase.SG.NOM table.SG.ADE / table.sg.gen on be-prs.3SG pretty.SG.NOM
'The vase on the table is pretty.'

2.2 The Complexity Principle (CP)

The Complexity Principle (Rohdenburg 1996, 2002, 2021; henceforth: 'the CP') aims to establish a direct relationship between the cognitive complexity of a grammatical construction and morphosyntactic choices made in connection with it. We adopt the CP as a theoretical framework since it (i) is cognitively grounded, (ii) generates

concrete, testable hypotheses and (iii) has been applied to phenomena broadly compatible with the alternations we look at. On the other hand, the CP was developed in English linguistics, and research based on it largely focuses on English. The transfer of CP-related principles and assumptions to Estonian – a language typologically remote from English – is probably not without its complications. However, the validity of the CP as a *principle* will have to depend on its cross-linguistic applicability or generality. Rohdenburg (1996: 151; also Mondorf 2009, Ch. 2) explains that the CP is an extension of theories proposed by Hawkins (e.g., 1990, 1992) but goes beyond Hawkins's focus on the processability of word-order alternations. It is formulated thus (Rohdenburg 1996: 151): "In the case of more or less explicit grammatical options the more explicit one(s) will tend to be favoured in cognitively more complex environments."

In a later re-statement (Rohdenburg 2002: 79), the CP is extended to 'grammatical (or lexico-grammatical) options'. Two fundamental questions come with the CP: (i) What makes a construction cognitively complex, and (ii) which grammatical variants qualify as cognitively more or less explicit (or transparent)? Concerning the first question, Rohdenburg (2021: 771) lists several circumstances that make a construction more complex, for instance, passivisation, subordinate clause negation, longer clauses or more complex noun phrases, as well as the low frequency of items (and, by extension, constructions). Based on examples like these, it is evident that complexity is a flexible and general concept that can apply at different levels. In many cases, it correlates with length because heavy constructions tend to be more challenging to store in working memory during parsing. In other cases, there need not even be a direct link to structural properties, for instance, when we assume that more frequent items are more deeply entrenched, easier to access, and therefore less complex. Finally, general notions of markedness also play a role, although Rohdenburg does not state this explicitly. If passives and negated subordinate clauses are classified as more complex, as in Rohdenburg (2021), then this is neither based on the length/weight nor exclusively on the frequency of such constructions but on the fact that an alternative unmarked (e.g. active or affirmative) construction exists. The broad scope of 'complexity' is part of its appeal, but there is also some potential for overextending the notion to contexts in which it does not truly hold. In other words: Its flexibility also makes the CP intrinsically vague and dependent on other assumptions – for instance, the notion of markedness that underlies several of the examples above has itself not gone uncontested (e.g. Haspelmath 2006). We would also expect such issues to be aggravated when we take Rohdenburg's CP beyond its original 'habitat' – the English language – to another, typologically different language like Estonian.

Rohdenburg (2021: 771) lists grammatical alternations regarded as more or less explicit, some of which we show in Table 1. Concerning this matter, Rohdenburg

(1996: 151–152; cf. 2021: 771) generally focuses on "formal contrasts involving the deletion (or addition) and the substitution of grammatical (or closed-class) elements", so that "the more explicit variant is generally represented by the bulkier element or construction". Explicitness (or transparency) thus correlates with more substantial morphological material. As we will argue below, this relationship between morphological substance and explicitness/transparency works well for the cases adduced by Rohdenburg throughout his work. However, here, too, there is a danger of assuming that the relationship is a universal one.

Mondorf (2009) studies the English comparative alternation in English (*heavier* vs. *more heavy*) in a CP-based framework. Her expectation that more complex adjective phrases correlate with the use of analytic comparatives rests on the assumptions that (i) this provides an early indication that a comparative follows and thus makes it easier to identify phrase structure, (ii) morphological parsing is simplified by encoding the degree-marking morpheme as a free item rather than an inflectional affix, and (iii) the phrase-initial marker *more* alerts the parser not only to expect a comparative, but also to expect a relatively complex AdjP, and thus to activate a greater processing capacity (2009: 6–7). Mondorf (2009: 6) suggests that *more*-support in English comparatives "can most likely be extended to other kinds of variation phenomena that draw on the synthetic-analytic distinction" because it "merely represents one instance of a presumably universal tendency". She calls this tendency *periphrasis-support* or *analytic support* and defines it thus:

> In cognitively more demanding environments which require an increased processing load, language users – when faced with the option between a synthetic and analytic variant – tend to compensate for the additional effort by resorting to the analytic form. (Mondorf 2009: 6)

However, it is immediately clear that this new principle needs to be viewed critically in the context of Estonian exterior locatives. Here, the analytic variant hinges upon a postposition, which means that only the second point made by Mondorf (whereby free morphemes are generally more explicit than affixes) applies. In contrast, her

Table 1: Grammatical alternations in English involving more or less explicit variants (from Rohdenburg 2021: 771).

Function	more explicit	less explicit
introduction of complement clauses	complementizer *that*	zero
complementation of mandative predicates	*should* + infinitive	subjunctive
verbal concepts as subject/object	*to*-infinitive	gerund
resumptive pronouns	yes	no
comparatives (cf. Mondorf 2009)	analytic	synthetic

first and third points do not. Moreover, as explained above, in the inflectional variants of Estonian exterior locative constructions it is not simply the headword in a complex Landmark phrase that gets inflected, but also its preceding modifiers. Thus, while the analytic variant of an Estonian exterior locative construction comes with a separate (and therefore, perhaps, maximally transparent) adposition, the inflections in a synthetic variant aid the parser even before the headword is reached. Mondorf's (2009) derivation of an analytic-support principle from Rohdenburg's CP works in English, where analyticity and anticipatory word order coincide. However, it is rather doubtful whether it can be transferred to Estonian.

2.3 Research questions and expectations

We treat the three semantic relations of lative, locative and separative as subordinate, more specific manifestations of a more general exterior locative concept. While they are characterised by different relations between Trajector and Landmark in terms of dynamics and direction, our working assumption still is that they are closely related enough to be governed by similar grammatical constraints. From a cognitive perspective, this kind of probabilistic uniformity is economical because it requires fewer 'rules'. Beyond this expectation of relative homogeneity, we have more specific expectations motivated by the CP. We formulate the following hypotheses:

Hypothesis 1: The variation of lative, locative and separative constructions (analytic vs. synthetic) will follow similar general principles because they are stored as subtypes of a single overarching exterior locative super-construction. In our quantitative analyses we expect the direction of individual effects tested under Hypothesis 2 to be the same for all three.

Hypothesis 2: The selection of an analytic (postpositional) over a synthetic (inflectional) realisation will be governed by the cognitive complexity of the construction at several levels, whereby cognitively more complex constructions favour the postpositional realisation. Specifically, we treat as more complex

a. Landmark phrases that precede the associated Trajector phrase,
b. Landmark lemmas that are compounds,
c. Landmark phrases that are relatively long, and
d. Landmark forms that occur relatively rarely in combination with the respective locative relation.

The first, more global assumption can hold even if the more specific predictions under the second hypothesis do not. In other words: Confirmation of Hypothesis 1 is

necessary for (but does not imply) the confirmation of Hypothesis 2, while confirmation of Hypothesis 2 implies confirmation of Hypothesis 1. We make predictions only about the directions of effects, not about their strengths. There is, at present, no theoretical reason for us to assume that one or the other semantic relation should be more strongly affected by any one of the factors.

3 Data and methodology

In this section, we discuss our data retrieval and downsampling procedures (3.1), the coding of the data (3.2), and the statistical methods of analysis (3.3). The essential statistical details documented in Section 3.3 may only interest some readers, and it should be possible to follow our discussion of visualised results in Section 4 even if large parts of Section 3.3 are skipped.

3.1 Data retrieval

Our data were extracted from the Estonian National Corpus (ENC 2017; 1.1 billion words, mainly web-based; Kallas and Koppel 2018) using SketchEngine (http://www.sketchengine.eu/). The corpus has been automatically tagged with the Estonian Filosoft part-of-speech tagset allowing for the automatic extraction of case forms and postpositions.

Subsampling was necessary because the manual annotation of occurrences is rather time-consuming (see Section 3.2). Due to the generally much lower frequency of the postpositional variant for all three constructions, we decided to take random samples of 1,000 cases for each of the three alternations, balanced by variant (with $n = 500$ synthetic and $n = 500$ analytic cases), which resulted in a total of $n = 3{,}000$ cases. Without this kind of balancing, the number of postpositional cases would have been too small for a statistical analysis involving the set of predictors we used. Table 2 documents the number of hits obtained for each of the three constructions and the steps we took in generating a manageable subsample. Since our approach has consequences for estimating the percentages in Section 4 below, we will briefly explain it here. Within the lative construction in Table 2, there were a total of $F = 959{,}515$ cases marked by the postposition *peale.* Inspecting a random sample of these, it took $f = 2{,}142$ cases before we arrived at the targeted number of $s = 500$ valid cases. The rejected 1,642 cases did carry the postposition *peale* but expressed a different semantic relation. False positives of this kind are often invariable – that is, they do not allow for the alternation between postposition and inflection. Our precision *Pr*

Table 2: Documentation of the sampling process.

Construction	*F*	*f*	*s*	*Pr = s/f*	$F_a = Pr * F$	p_a
Lative						
Allative	19,187,296	3,017	500	0.166	3,179,863	0.934
peale	959,515	2,142	500	0.233	223,976	0.066
Locative						
Adessive	30,661,120	5,148	500	0.097	2,977,964	0.968
peal	241,263	1,210	500	0.413	99,695	0.032
Separative						
Ablative	2,675,044	1,745	500	0.287	766,488	0.906
pealt	138,049	872	500	0.573	79,157	0.094

for *peale* was, therefore, 500/2,142 = 0.233; thus, we concluded that only 23.3 % of instances of *peale* in the ENC express the lative relation, and this figure was used to calculate the adjusted absolute frequency F_a of *peale*. We finally arrived at adjusted proportions (p_a) of postpositional and inflectional variants based on the pair of adjusted values within each of the three constructions. At this point, suffice it to say that these extrapolations are essential to our statistical analysis, as will be explained in Section 3.2.

From the overall frequencies of the three exterior locative cases and the corresponding adpositions in Table 2, it can be seen that the most frequent case in the trio is the adessive, followed by allative and ablative. The most frequent postposition is *peale*. The adessive has the lowest value for precision, since this case affix has other essential functions to fulfil in the language (e.g., possessor in the possessive construction; see Appendix A).

3.2 Annotation

All occurrences were coded by the first author for nine variables, as shown in Table 3. The outcome variable is POSTPOS, with *case* as the reference level and *postposition* predicted by the regression model (see below). The other variables are discussed below.

The predictor POSITION describes the arrangement of the Landmark phrase and Trajector relative to each other, with 'pre' standing for Landmarks preceding their Trajector (example 6) and 'post' for the inverse pattern (example 7). We treat as more complex the word order where Landmark phrases precede the associated Trajector phrase, and we predict that a Landmark preceding the Trajector favours postpositions (Hypothesis 2a).

Table 3: Definition of variables.

Variable	Category	Scale/levels (reference level stated first for categorical variables)
POSTPOS	outcome	*case* *postposition*
POSITION	fixed	*post* *pre*
COMPLEXITY	fixed	*simple* *compound*
LENGTH	fixed	$\log_2$-length (in syllables) of Landmark phrase, centred on the value 2 (~4 syllables).
FREQUENCY	fixed	$\log_2$-frequency (raw) of Landmark form in association with the respective semantic relation, centred on the value 9 (~512 occurrences).
FUNCTION	fixed	*adverbial* *modifier*
VERB_LEMMA	random	224 levels (lative) 279 levels (locative) 252 levels (separative)
LM_LEMMA	random	592 levels (lative) 438 levels (locative) 528 levels (separative)
SEM_REL	filter	*lat* (= lative) *loc* (= locative) *sep* (= separative)

(6) **Põrandal** lebab Vogue
floor.SG.ADE lie-PRS.3SG Vogue.SG.NOM
'A Vogue is lying **on the floor**.'
(Lit. '**On the floor** is lying a Vogue.')

(7) Maja asus **mäekalda** **peal.**
house.SG.NOM be.located-PST.3SG mountain.slope.SG.GEN on
'The house was located **on a mountain slope**.'

In terms of topic-comment structure, Landmark preceding Trajector is more marked.[1] The Trajector is considered to be the most prominent participant in locative expressions (Langacker 2008: 70), but it is in the relationship between Trajector and

1 The assumption that Given precedes New has been questioned in the typological literature for some languages, and it depends on how "given" and "new" are defined; e.g., Mithun (1987) refers to "newsworthiness" rather than "new". We are grateful to the anonymous reviewer for pointing this out.

Landmark that new information is found. It stands to reason, therefore, that the preferred or canonical word order is such that the Trajector phrase precedes the Landmark phrase. Although Estonian is typically considered a language with a relatively free word order (Lindström 2005: 10), any (functionally oriented) linguist will agree that nothing is ever completely 'free'. In Estonian, as in many other European languages, it is common to begin a clause with the information already known to the speaker/listener and provide new information at the end of a clause. According to Estonian reference grammars (e.g., Erelt 2003: 93–94), Estonian has two basic syntactic patterns of clauses: normal and inverted. The basic word order in the normal clause is SVA, where the subject (S) is morphologically unmarked, the verb (V) agrees with the subject in person and number, and the verb may be followed, for example, by a subject predicative and/or a locational adverb (A) (cf. example 8).

(8) Jaan on toas.
Jaan.SG.NOM be-PRS.3SG room.SG.INE
'Jaan is in the room.'

In the "inverted clauses", it is not the subject that comes first in the clause, as in normal clauses, but an adverbial (A) or an oblique object (Obl) expressing location, time, possessor or experiencer (Erelt 2003: 93; cf. example 9). Lindström (2005: 10) refers to these clauses as specific clause types in which the word order sequence is conventionalised in Estonian. Such conventionalised word order sequences make it easier for the listener/reader to understand the clause since information structure plays a role in these clauses. What is already known comes at the beginning, and what is new at the end of the clause (Lindström 2005: 10).

(9) Aias on lilled.
garden.SG.INE be-PRS.3PL flower.PL.NOM
'There are flowers in the garden.'
(lit. 'In the garden are flowers.')

Given that the adpositional variant is more explicit and specific than the case variant, it should be used at the end of clauses where new information is provided in Estonian. The case variant is predicted to be used at the beginning of a clause because it is shorter than the adpositional variant and less specific.

For the predictor LENGTH, the number of syllables in the Landmark phrase was counted. The binary logarithm (base = 2) was applied, and values were centred on the whole-number value closest to the mean, which happened to be 2 (corresponding to a length of 4 syllables). We regard relatively long Landmark phrases as more complex and therefore predict that they should favour postpositions (Hypothesis 2c). Length is one of the most crucial variables in numerous studies on various syntactic alternation phenomena (e.g., Hawkins 1994; Wasow 1997; Arnold et al. 2000; Wasow 2002;

Wasow and Arnold 2003; Hawkins 2004). A critical methodological question concerns whether to include the adpositions *peale, peal, pealt* in the count as a separate word. 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 possessum phrase of an *of*-genitive were not included because they provide a natural imbalance and skew the results. For similar reasons, it was decided not to include the postpositions *peale, peal, pealt* in the counts. Hence, Landmark phrases such as *laual* [laud + ADE] and *laua peal* [laud + GEN peal] were considered disyllabic. It should be noted that with adjectival attributes, there is an agreeing case marker on the attribute, e.g., *i-lu-sa-le lau-a-le* (7 syllables) versus *i-lu-sa lau-a peale* (5 syllables), making the former phrase longer and hence more complex than the latter.

The predictor COMPLEXITY takes the reference value *simple* if the Landmark lemma is not a compound (e.g., *laud* 'table' in example 10) and the value *compound* if it is (e.g., *kirjutuslaud* 'writing desk' in example 11). We treat compound Landmarks as more complex and, following the Complexity Principle, we predict that compound Landmarks favour postpositions (Hypothesis 2b).

(10) **Laua** **peal** aga seisis mitut
table.SG.GEN on but stand-PRS.3SG many.SG.PRT
arvutit meenutav aparaat.
computer.SG.PRT remind-PRS.PCPL machine.SG.NOM
'But a machine resembling many computers was standing **on the table**.'

(11) Trengi sõrmed alustavad tantsu **kirjutuslaual**.
Treng.SG.GEN finger.PL.NOM begin-PRS.3PL dance.SG.PRT writing.desk.SG.ADE
'Treng's fingers begin a dance **on the writing desk**.'

FREQUENCY denotes the overall raw frequency of a Landmark form in combination with the respective semantic relation (lative, locative or separative) in the entire ENC. This, too, was logged (base = 2) and centred on the whole number closest to the mean, which happened to be nine (corresponding to an absolute f = 512). We treat less frequent combinations as more complex and predict that they will favour postpositions (Hypothesis 2d). This is based on the assumption that form-function combinations of lower frequency will be less entrenched, harder to access and therefore more complex, not in a formal but in a cognitive way.

The variable FUNCTION denotes the grammatical function of the Landmark phrase: *adverbial* means that the Landmark is a sentence constituent in its own right, while *modifier* means that the Landmark modifies the Trajector within the same constituent. For example, both the adessive and the adpositional realisation with *peal* can fulfil two syntactic functions in a clause – that of an adverbial, as *õue peal* 'on the yard' in example 12, or an adverbial modifier as *merel* 'on the sea' in

example 13. The variable has two levels: 'adverbial' and 'modifier'. Although it may be presumed that the modifier function is more complex since it involves embedded modification, i.e. it is a type of environmental complexity that the CP addresses and hence the more explicit, postpositional variant should be more likely, we have no solid prior expectation as to which of them will favour the use of a postposition.

(12) Kuule, sul kasvab **õue** **peal**
listen-IMP you.ADE grow-PRS.3SG yard.SG.GEN on
üks suur puu?
one.SG.NOM big.SG.NOM tree.SG.NOM
'Listen, you have a big tree growing **in the yard**, right?'
(lit. 'Listen, you have a big tree growing on the yard?)

(13) […] kaluritele ja teistele **merel** töötavatele
fisherman.PL.ALL and other.PL.ALL sea.SG.ADE working.PL.ALL
inimestele […]
person.PL.ALL
'[…] to fishermen and other people working **on the sea** […]'

The two self-explanatory variables VERB_LEMMA and LM_LEMMA were used as group variables in the random-effects part of our models (see below) and are not documented in detail, apart from the number of levels they take for each construction. Finally, SEM_REL is classified as a 'filter' variable in Table 1: Its only function is to partition the data into three subsets, which were then used in the respective models, one for each construction.

3.3 Statistical modelling and visualisation

We worked in the R-environment (R Core Team 2019), using RStudio (RStudio Team 2009–2019). For each of the three constructions (lative, locative, separative), we fitted a Bayesian mixed-effects logistic regression model using the R-package 'brms' (Bürkner 2020), which is based on Stan (Stan Development Team 2019). Each model drew on a subset of the data defined by the variable SEM_REL (see Table 1) and was specified as follows:

POSTPOS ~ COMPLEXITY + FUNCTION + LENGTH + FREQUENCY + POSITION
+ (1 | VERB_LEMMA)
+ (1 | LM_LEMMA)

No interactions were included in the model, as we have no theoretical reason to assume any moderating effects among predictors. No random slopes were specified for the group variables VERB_LEMMA and LM_LEMMA; thus, we decided against a maximised random part, as Barr et al. (2013) recommended. Weakly regularizing priors were specified for the fixed part, and given the estimated coefficients (which fell squarely within the prior distribution), we dispensed with a sensitivity analysis of alternative priors. Somewhat more restrictive priors were used for the random part. The model syntax and the priors were the same for all three models, but the number of levels for both group variables (VERB_LEMMA and LM_LEMMA) differs between them, as documented in Table 2 and Appendix B.

Models were run with four chains of n = 3,000 iterations, each including a warmup phase of 1,000 iterations. The total number of posterior samples was thus n = 8,000. The convergence of chains was indicated by the R-hat diagnostic taking the value of 1.00 for all parameters in the model. A basic summary of model coefficients and priors is given in the appendix, but also see our discussion of Figure 1 below. A more comprehensive summary can be found in the online repository (see data availability statement). No model comparison or model selection process was conducted since it was considered essential to retain all theoretically important predictors in the model, irrespective of their estimated effects (cf. Heinze et al. 2018); we thus adopted the notion of the 'deductive model' as proposed by Tizón-Couto and Lorenz (2021).

Visualisations were done with functions in the R-package 'lattice' (Sarkar 2018). Plots show the median posterior probability (in %) of the postpositional variant under the respective condition and 50 % and 90 % percentile-based posterior uncertainty intervals. Such intervals will occasionally be reported in the text. For instance, in the third panel of Figure 2, the percentage of postpositions for the lative in connection with short LMs can be given as '15.2 % [11.7; 19.3]', which means 'a median posterior percentage estimate of 15.2 % with a 90 % uncertainty interval extending from 11.7 % to 19.3 %'. We also display the contrast between conditions (e.g., short Landmark phrases vs. long Landmark phrases) as an estimated difference in absolute percentage points with the respective uncertainty intervals. Comparing such intervals to the reference value of zero (= 'no difference') provides additional evidence for interpreting an effect beyond the effect size itself. The difference between short and long LM phrases was estimated with values of ±1 for the centred predictor LENGTH, which is the whole-number value closest to one standard deviation (SD) of this predictor across the data. The same approach was taken for FREQUENCY, in this case, predictor values of ±4 corresponded most closely to ±1 SD. Factors of no immediate interest in a given scenario are held constant. The exact routine is made transparent in the online material (see data availability statement below). For categorical variables like FUNCTION or POSITION, it may seem counterintuitive to assume

normal (or average) values since, in reality, they take categorical values at any one time. However, this approach allows us to target specific effects and thus make results more accessible.

As explained in Section 3.1, the share of postpositional realisations in the ENC is much lower than the share of 50 % in our balanced dataset. For instance, we expect the actual proportion of *peale* within the targeted semantic context (lative) to be only 0.066 (or 6.6 %), not 0.5 (or 50 %) as in our data. In order to model the actual expected share of postpositions more realistically, we added a corrective intercept to our estimates, calculated as the log-odds of the 'real' proportion of the outcome variant.[2] The values we used were $\log(0.0658/(1 - 0.0658)) = -2.65$ for the lative, $\log(0.0324/(1 - 0.0324)) = -3.40$ for the locative, and $\log(0.0936/(1 - 0.0936)) = -2.27$ for the separative (see Table 2).

4 Results

The results part of this article breaks down into five components. We begin with a short discussion of model coefficients (4.1). While relatively abstract, this is useful as it indicates the direction of individual effects and shows whether or not our hypotheses are generally supported. In the second part (4.2), we isolate the four effects attached to our Hypotheses 2a–d (see 2.3 above). Thirdly (4.3), we show outcomes for specific combinations of predictor values to give the reader an idea as to the actual range of percentage values our model predicts. In Section 4.4, we correlate specific conditions across the three locative constructions to test our general first hypothesis, which stated that lative, locative and separative should be similarly affected by the included predictors. Finally, in Section 4.5, we will say a few words about the effects the grammatical status (adverbial vs. modifier) of the Landmark has on the outcome, even though we did not theorise about this predictor or include it in our hypotheses.

4.1 Model coefficients

We first turn to a short discussion of the coefficients in the fixed part of the statistical model, shown in Figure 1 (also see Appendix B). Four predictors appear in black. These correspond to the four hypotheses formulated as 2a–d in Section 2.3 above. The

2 Our approach is perhaps less common in quantitative corpus linguistics but quite well-known in other disciplines dealing with rare events, e.g. medical research. For instance, the number of people contracting a particular disease may be so small that a random sample would not capture enough cases to make meaningful statements about those infected. Therefore, a sizeable number of patients is balanced against a sample of controls.

direction of an expected effect is indicated with the letter 'H' in the margin of the plotting region: If the respective predictor has the expected effect, its coefficient should be positioned between the vertical reference line at zero and this symbol. The fifth predictor, FUNCTION, appears in grey and without the symbol 'H' because it does not come with a particular expectation but was included as a control. It will, therefore, not feature in the main discussion, although we will briefly return to it in Section 4.5. Note that the expected direction of the effect is negative for the predictor FREQUENCY, while for the other three, it is positive. Note further that we use stricter uncertainty intervals of 50 % and 95 % in this plot, while in the percentage plots below, we use 50 % and 90 % intervals, as discussed in Section 3.3.

Results transformed into percentages will be discussed in more detail below. What we can see at a glance in Figure 1, however, is that for the predictors COMPLEXITY and LENGTH, the effect we find in the data runs counter to our expectations (cf. Hypotheses 2b and c): Compound LMs as well as LMs of greater length *dis*favour the selection of the postpositional variant. LMs of higher frequency (in connection with the respective alternations) also disfavour the selection of the postpositional variant, but this is a pattern we expected in connection with Hypothesis 2d. The actual effects of LENGTH and FREQUENCY are larger than the figure suggests since we assume that it is not a unit change but a change of 1 SD that should be treated as a realistic change in the predictor value (see discussion in Section 3.3 above). The remaining theoretically motivated predictor, POSITION, does not affect the variation of the lative; for the other two semantic categories, it tends to support our Hypothesis 2a (with a correlation between postpositions and LMs that precede the Trajector). However, the effect is not very pronounced and comes with a higher degree of uncertainty, with the 95 % intervals cutting across the zero reference value.[3] While

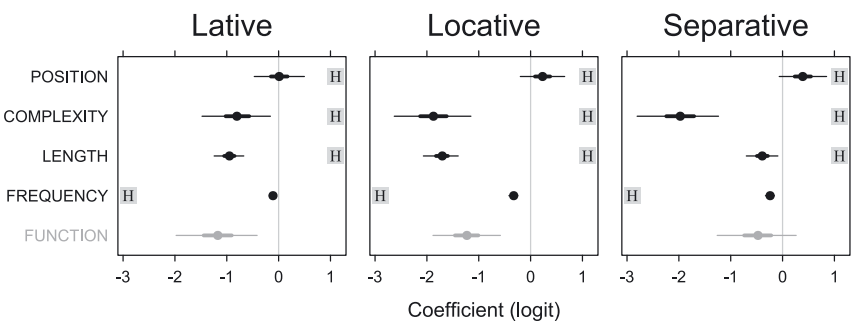

Figure 1: Coefficient estimates CC BY.

3 The effect would likely be non-significant in a frequentist (non-Bayesian) model of the same structure.

effect sizes are not equal, the general pattern is the same for all three semantic categories, which provides a first pointer to our first broad hypothesis concerning the uniformity of mechanisms of variation across all three semantic categories (also see 4.4 below). In the next section, we will unfold these relatively general and abstract findings into predicted percentages of the postpositional variant.

4.2 Average probabilities by individual factors

To facilitate the direct comparison of the predictors mainly of interest, all four are combined in Figure 2. The panel corresponding to each predictor breaks down into two parts: The bottom part shows estimated percentages of the postpositional variant under the two contrasting conditions (e.g., simple vs. compound LMs in the second panel) for the three semantic categories, while the top part shows the estimated absolute percentage-point difference between conditions. Both parts of each panel show 50 % and 90 % posterior uncertainty intervals. Despite the traditional criticism that lines should not be used to connect categorically different points, we still use them because they have considerable advantages for directly comparing the patterns (cf. Sönning forthcoming). The colour scheme is adjusted to make transparent whether or not our hypotheses are confirmed: In the bottom part of each panel, black points above grey points signal that the data support the respective hypothesis; this pattern would then result in points positioned above the grey reference line in the corresponding plot on top.

When we average across the two contrasting conditions, irrespective of the panel we look at, it can be seen that separative constructions are generally most likely to be realised with a postposition (18.1 %), followed by locatives (14.8 %), while postpositions are least likely with the lative (9.9 %). This pattern is most clearly visible in the bottom part of panel one (POSITION) because it holds true under both conditions and because the difference between LMs that precede and those that follow the Trajector is small. Secondly, each predictor's effect has the same direction across all three constructions, although it differs in magnitude, as discussed in more detail below. This can be seen at a glance in the top panels, where all values for a given predictor are either above or below the reference line.

For separative meanings, an LM that precedes the Trajector is more likely to take a postposition than cases in which the LM follows the Trajector (+3.0 [0.1; 6.1]). This effect is in the expected direction, as we regarded the pattern LM → Trajector as the marked sequence that would require a more explicit (postpositional) signal. However, the effect is relatively small – for the locative, it must be called negligible, and for the lative it is virtually non-existent, with +1.6 [−0.9; 4.1] and +0.1 [−2.3; 2.4], respectively. There is thus only minimal support for our Hypothesis 2a.

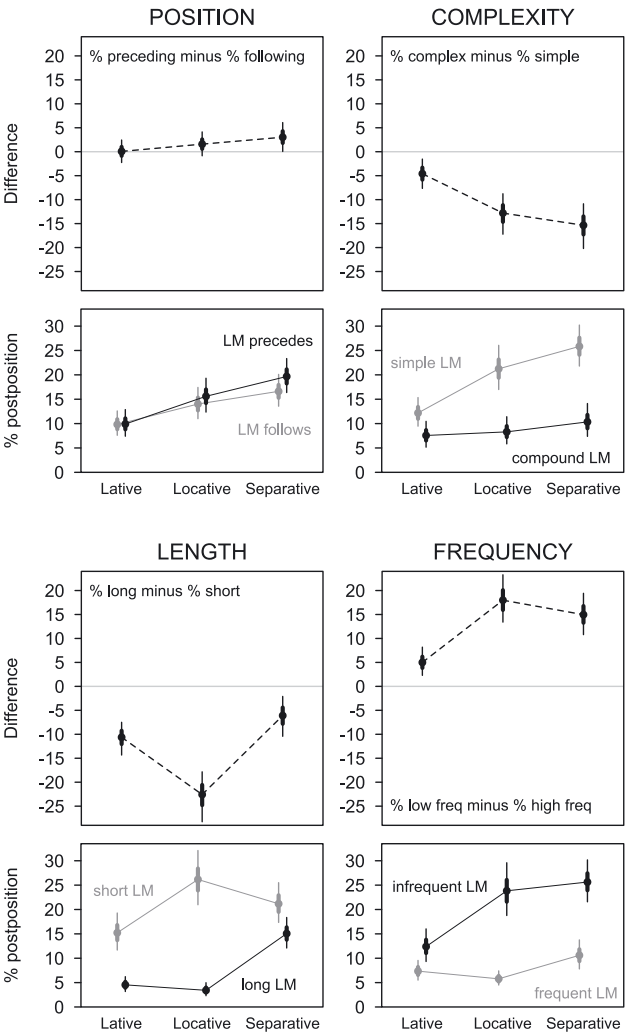

Figure 2: Posterior probabilities and contrasts by individual factors CC BY.

Regarding our second main predictor, COMPLEXITY, constructions with compound Landmarks are much less likely to be postpositional than inflectional in the separative (–15.3 % [–20.2; –10.9]) and the locative (–12.8 % [–17.2; –8.8]); in the lative, the effect is less pronounced but still very much visible – (–4.6 % [–7.7; –1.5]). These findings contradict expectations formulated as Hypothesis 2b: There is no evidence

that compound LMs favour the postpositional variant, and the exact opposite pattern prevails.

The third panel of Figure 2 shows that the length of the LM substantially influences the selection of constructional variants. Longer LM phrases are less likely to combine with postpositions than constructions with shorter LM phrases. The general direction of this effect is the same for all three alternations, and it goes directly against our Hypothesis 2c: Since longer LMs involve a greater processing load, we argued that they should be more likely to combine with the more explicit (namely postpositional) markers. In this case, the effect is most pronounced with the locative (–22.6 % [–28.3; –17.8]), followed by the lative (–10.6 % [–14.3; –7.5]), while the separative is least affected (–6.1 % [–10.4; –2.1]).

Finally, constructions with LM forms that occur less frequently with the respective locative function are generally more likely to take postpositions than forms that are more frequently used. This effect is most pronounced with the locative (+18.0 [13.4; 23.3]), followed by the separative (+15.0 [10.8; 19.4]), while the lative is least affected (+5.0 [2.3; 8.2]). Thus, FREQUENCY is the only predictor that genuinely supports its associated hypothesis (2d).

There are two central conclusions from the above. Firstly, results broadly agree with our expectation that constructional variation in the categories (lative, locative, separative) should be conditioned along similar lines by our main predictors. The effect of any one predictor may differ in magnitude between semantic categories. However, it always carries the same sign – or, in other words, it never points in the opposite direction. We therefore regard our first hypothesis as confirmed. A subsidiary finding is that there is no obvious ranking of the three semantic relations based on their responsiveness to conditioning factors: Apart from the uniform direction of effects in Figure 2, the patterns we see in the top parts of the four panels are all different – except perhaps that the lative is least affected by three out the four predictors: POSITION, COMPLEXITY and FREQUENCY. The second major finding is that of our four specific hypotheses, only the one relating to frequency (hypothesis 2d) is strongly supported.

4.3 Percentages for specific factor combinations

While the preceding section isolated the effects of the four predictors associated with our Hypotheses 2a–d, we will look at specific combinations of factor settings in the following paragraphs. By not averaging across specific conditions, we will gain insights concerning the overall range of percentages predicted for each semantic category and, thus, the reality underlying the more generalised results presented

above. The only predictor not considered in this approach is FUNCTION; this we control for since it does not feature in our main assumptions.

Figure 3 shows sixteen specific conditions for each of the three constructions, arranged in descending order based on the estimated median percentages of postpositional variants. Since individual effects point in the same direction for all three alternations, the ranking is bound to be similar if, of course, not quite the same. To facilitate processing, no uncertainty intervals are shown in this plot. Conditions are identified by four columns of black and white circles on the right, corresponding to the four predictor variables POSITION, COMPLEXITY, LENGTH and FREQUENCY. Circles are labelled '1' (LM precedes TR) versus '2' (LM follows TR) for POSITION, 'S' (simple/non-compound) versus 'C' (complex/compound) for COMPLEXITY, 'S' (short) versus 'L' (long) for LENGTH: and 'H' (high) versus 'L' (low) for FREQUENCY. The colour coding again reflects our expectations: According to our Hypotheses 2a–d, black dots should be nearer the top and white dots should be nearer the bottom in each panel.

For the locative and lative, the top six and bottom six conditions are the same; furthermore, the top two and bottom two conditions are the same between all three semantic relations. These findings in themselves already indicate that there is a considerable degree of similarity between the different locative relations regarding their responses to the factors we included.

If we ignore POSITION, which has only a minor effect (cf. Figure 2), simple, short and infrequent LMs are most strongly in favour of postpositional variants: If averaged across the two respective conditions for POSITION, the expected percentages are 22.5 % in the lative, 40.5 % in the separative and 56.4 % in the locative. Naturally, the lowest values are found with LMs that are compound, long and frequent, with average percentages of 3.8 % (separative), 2.1 % (lative) and 0.2 % (locative), respectively. The lative displays the most moderate range of values, which agrees with the finding that it was also the least affected under most circumstances in Section 4.2. Separative and locative, conversely, are characterised by more variation between conditions, the latter with two rather extreme high values. Note that the specific scenarios pictured here are not necessarily frequent in natural language production in Estonian; the sometimes high percentages shown here do not conflict with the generally much lower frequencies of postpositions in the corpus.

4.4 The uniformity of effects across constructions

It was evident from Figures 1 and 2 that effects, if not equal in size and if not consistently ranked in the same way across semantic functions, nevertheless point in the same direction and that estimates for specific conditions should therefore be correlated across semantic types. We explore this finding more formally in Figure 4,

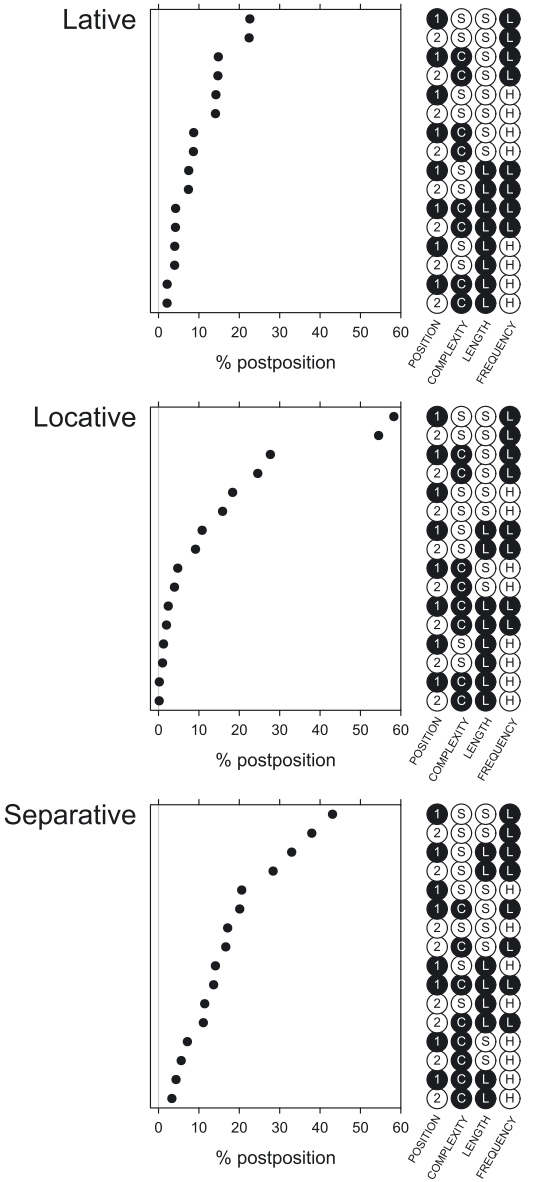

Figure 3: Ranked median percentages for individual factor combinations by construction.

which ranks the n = 16 specific conditions defined by binary values of POSITION, COMPLEXITY, LENGTH and FREQUENCY according to the average predicted frequency of the postpositional variant across the three semantic relations, in ascending order from left to right.

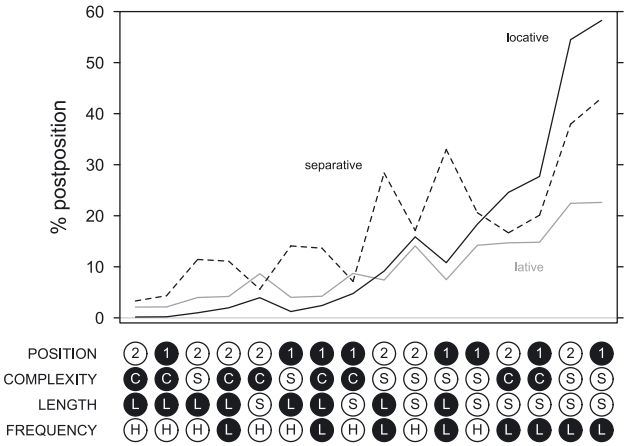

Figure 4: Assessing the uniformity of effects across constructions.

A more or less regular upward slope can be observed in all three relations – more regular for the locative and the lative, less so for the separative. In other words: Conditions that result in a relatively high share of postpositional realisations on average tend to do so for all three semantic categories, which is confirmation of our most general first hypothesis (see Section 1.2). As mentioned above, this finding is not new at this point – it is merely shown from a different perspective. The (sometimes striking) contrasts between the separative and the other two categories at individual steps in the left-to-right progression of Figure 4 are due to different effect sizes, not to differences in the direction of effects (see Figure 1 and Appendix B).[4]

Table 4 further confirms the broad correlation of the behaviours of lative, locative and separative under the four conditioning factors we consider here. We use one linear and one rank-based correlation coefficient.

While all three possible correlations are statistically significant, we can also see that, by far the strongest correlation is the one between the lative and the locative. The separative does not correlate as well, which agrees with its somewhat more erratic pattern in Figure 4. Our first hypothesis only addressed the expected general alignment of the three alternations regarding their responses to the different predictor variables; we had no theory to motivate why any of them should diverge from the average pattern.

4 The 'reversals' in the separative may initially seem odd and counterintuitive, but they can be traced by inspecting how exactly conditions change between the respective 'steps' of the plot. Readers who wish to understand this in detail are advised to look out for the small effect of LENGTH on variation in the separative (cf. Figure 1 and Appendix B).

Table 4: Correlation of median percentages between constructions .

Correlation	Pearson		Spearman	
	r	*p*	Rho	*p*
lative ~ locative	0.95	0.000	0.98	0.000
lative ~ separative	0.74	0.001	0.71	0.003
locative ~ separative	0.82	0.000	0.83	0.000

4.5 Short comments on the grammatical status of Landmarks

This section highlights the difference between LMs that function as adverbials at the sentence level and those constructed as modifiers within a noun phrase. The discussion will be relatively brief: The predictor FUNCTION, while included in our statistical model, is detached from our hypotheses – that is, we saw no reason to expect that either adverbials or modifiers should favour postpositional variants. However, find an effect we did, and while its theoretical implications remain opaque for now, we decided to include it in our presentation.

Figure 5 shows that LMs functioning as adverbials correlate with higher proportions of postpositions, with a percentage-point difference (relative to the modifier use) of +6.7 % [3.3; 9.8], +8.4 % [5.0; 11.8] and +3.7 % [–1.3; 8.3] for the lative, locative and separative, respectively. Although a higher degree of uncertainty comes with the result for the separative, this is a regular finding, and the differences are substantial. As to accounting for this effect, we can at present only make the informal observation that the co-occurrence of higher-order syntactic elements (i.e., adverbials as sentence-level constituents, not modifiers as phrase-level constituents) with higher-order grammatical markers (i.e., free postpositions, not bound case affixes) seems intuitively plausible.

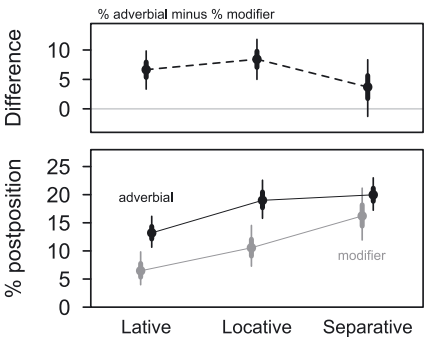

Figure 5: The effect of the predictor FUNCTION .

5 Discussion

Our focus was testing the applicability of the Complexity Principle (CP) to three morphosyntactic alternations in Estonian that involve marking a locative function with either a postposition or a case affix. The CP states that more explicit grammatical alternatives (in our case, the postpositional variant) tend to be preferred in cognitively more complex environments (Mondorf 2009; Rohdenburg 1996, 2002, 2021). Crucially, the CP treats bulkier (or morphologically more substantial) elements as more explicit, presenting a seemingly clear-cut case when it comes to the alternation between postpositions and (less bulky) inflections. We looked at cognitive complexity from different perspectives, inspecting the compound/non-compound status of the Landmark lemma, and the length, frequency and position of the Landmark phrase (i.e., the phrase expressed with either a postposition or a case affix). Our paper extends the CP-framework beyond studying explicitly syntactic alternatives and word order variants (cf. Hawkins 1990, 1992, 1994), instead focussing on morphosyntactic units that differ only in morphological weight (or independence) but are invariably placed at the end of the governing term.

We treated the three semantic relations of lative, locative and separative as manifestations of a more general exterior locative concept. This assumption motivated our first hypothesis (Hypothesis 1), which predicts that similar constraints should govern these constructions and that the direction of the explored individual effects should therefore be uniform across the three of them. This hypothesis was confirmed: Although the magnitude of an effect associated with a particular predictor varies across constructions (and is negligible in a few cases), we never see the reversal of an effect between any pair of locative relations. On average, conditions that result in relatively high (or low) shares of postpositional realisations tend to do so for all three semantic relations. We further observed that the lative and the locative are most closely aligned, while the separative correlates less well. Although this slight irregularity does not call the general similarity of the constructions into doubt, it raises the question of why the separative should be set apart. We had no prior expectations of this sort and are therefore not in a position to offer a theoretically motivated explanation. One observation is that the separative function is less frequently expressed in Estonian than the other two. It is conceivable that the higher frequencies of lative and locative lead to mutual reinforcement and deeper entrenchment of constraints, which extends to the separative to a slightly lesser extent. This construction would then be a somewhat less tightly integrated member of the set that forms the locative super-construction we posited. An interpretation along these lines is broadly in harmony with usage-based views of language variation (cf. Bybee 2013, *inter alia*). A second explanation is that, compared to the allative and

adessive cases, the ablative is characterised by a less polysemous array of functions (see Appendix A). Therefore, the pressure to regulate the structure of the separative construction (to which it belongs) in line with the locative super-construction may be lower. We must explicitly state, however, that both interpretations can be no more than points of departure for future research – for one, because the separative differs not categorically but by degrees; for another, because these suggestions are made post-hoc, with little theoretical underpinning.

For Hypothesis 2, we formulated a set of four different hypotheses (a–d) linked to the four predictors in our study: POSITION, COMPLEXITY, LENGTH and FREQUENCY. Out of the four, we find full support for Hypothesis 2d: As predicted, Landmark forms that occur less frequently with the respective locative function are generally more likely to take postpositions than forms that are more frequently used with the respective semantic relations. This result aligns with patterns known from linguistic typology more generally. It has been shown that a rare combination of case relation and Landmark gets a more "marked" expression (e.g., Aristar 1997; Comrie 1986). We found minimal support for Hypothesis 2a (related to POSITION): Landmark phrases that precede the associated Trajector phrase prefer the postpositional variant. The effect is small – but demonstrably different from zero – for separative meanings; for the locative, it is no more than a tendency; and for the lative it is virtually non-existent.

Neither of our Hypotheses 2b and 2c found confirmation. Hypothesis 2b stated that, due to their greater complexity, constructions with compound Landmarks should be more likely to be realised with postpositions; according to Hypothesis 2c, longer Landmark phrases were expected to combine more often with the postpositional variants. Not only is there no evidence in support of these two hypotheses, but the precisely opposite patterns are statistically relatively robust: Constructions with compound Landmarks are less likely to be postpositional than inflectional, and the same pattern is found for constructions involving longer Landmark phrases as opposed to ones with shorter Landmark phrases. To account for the mismatch between some of our results and the expectations we framed based on the CP, we reconsidered our assumptions regarding the predictors COMPLEXITY and LENGTH from three perspectives, the first two of which involve the following questions: (i) Did we make the correct assumptions about what is and is not complex in these two dimensions, and (ii) did we make the correct assumptions about what constitutes more or less explicit realisations? These two points need to be answered in an Estonian language context, and they directly raise the third, and perhaps most fundamental, question: (iii) Does the CP's failure to account for phenomena in Estonian mean that we must reject it as a general principle with cross-linguistic validity?

We will address the first two questions for both of our 'problematic' predictors in turn and make suggestions concerning alternative explanations. In doing so, and

particularly when turning to the third question, we will consider whether the CP takes a view of explicitness (or transparency) that is too narrow and makes it unsuitable as a framework for languages that are typologically remote from English, perhaps because they are highly inflectional, like Estonian.

Concerning the predictor COMPLEXITY, one could argue that compound LMs are, in fact, no more complex than simple ones precisely because they have been lexicalised and stored as single items. Giving them a categorically different status might, therefore, not be warranted. Any effect of greater lemma length (as a by-product of compositionality) would then be captured by our predictor LENGTH at the phrase level. However, this reasoning still does not explain why the pattern found for COMPLEXITY should not simply be neutral (or random) but the inverse of what we expected and fairly substantial at that.

Concerning the predictor LENGTH, 'longer' would seem to correspond to 'more complex', in the sense of 'more difficult to store in working memory' and therefore needing more structural support. The view that 'analytic' generally corresponds to 'more explicit', as implied in Mondorf's (2009) notion of 'analytic support', conflicts particularly with our findings for LENGTH. Nevertheless, why should constructions that are more difficult to parse due to their length strongly favour the less substantial, bound inflectional case affix rather than the more substantial, free postposition? We would argue that long phrases are not necessarily more difficult to parse in Estonian. Let us take a step back to Rohdenburg's original thoughts on the CP (see Section 2.2) and the inspiration drawn from the work of Hawkins. In this view, explicitness is very much about enabling an early recognition of phrase structures (and boundaries) and, thus, the immediate constituents of a sentence (cf. Hawkins 1990, 1992; Rohdenburg 1996). It does, therefore, not suffice to simply make a marker more prominent if, so to speak, it is unalterably in the 'wrong' (i.e., 'late') position. Notably, in the context of locative constructions in Estonian, it could even be argued that case affixes allow a slightly more speedy recognition of structure and function (because they are more explicit not in substance but in position): The affix is sometimes directly attached to the modified elements within the phrase (e.g., as in *suure**l** valge**l** laua**l***), and we effectively get early marking in constructions of this kind. As discussed above, the corresponding postposition consists of two components (e.g., *suure valge laua peal*), so it effectively generates more morphological material and shifts the marker (and thus the node for constituent recognition) even further to the right. It could also be argued that language economy (see, for instance, Sinclair 1991) counteracts the CP in our narrow sense: Unless there are reasons to select a heavier structure, language users will tend towards brevity. Due to the invariable post-position (relative to the LM) of both the affix and the adposition in the exterior locative, processability is lower if the analytic form is

selected. Against this argument, it could be said that postpositions are generally less ambiguous: Like the case affixes, they have a range of functions other than expressing the locative relations under investigation here, but as shown in Table 2, the proportion of postpositions that had to be manually excluded is smaller than the corresponding proportion of case-marked occurrences. On a different level, therefore, postpositional realisations are more explicit and easier to process because they are less ambiguous.[5] The question as to whether postpositional variants are more explicit will receive a different answer depending on whether we take a semantic or processing-based perspective, but there is evidence that inflection does, in fact, make it easier to process an exterior locative construction. In the English language, on the other hand, we get a somewhat different picture due to the absence of inflectional endings, which is a likely cause of the rare occurrence of postpositions.[6]

From our analysis, it seems that the Complexity Principle takes too narrow a view of what is explicit (or transparent) in a grammatical construction. In particular, the notion that explicitness is directly linked to morphological substance seems problematic. In English, which we can safely treat as the 'homeground' of the CP, morphological bulkiness typically correlates with an early marking (and thus recognition) of constituent structures, at least in the constructions successfully addressed with the principle (cf. Section 2.2). In other words: It is probably not the bulkiness of a marker that improves processability but its syntactic position. The notion that cognitively complex constructions tend to be marked more explicitly remains attractive and plausible. However, the equation of explicitness with morphological bulk in the CP may well be a fallacy resulting from the unique properties of the English language. If framed in this way, explicitness does not help to account for the variation of our Estonian constructions and will probably also fail to capture other phenomena in other languages. In particular, the derived principle of 'analytic support' (Mondorf 2009) is probably conditional upon other factors: In English, it only holds true because the respective free morphemes (as in the *of*-genitive or the *more*-comparative) precede the phrases to which they relate, and because there are no phrase-internal inflections. In Estonian, on the other hand, we are looking at postpositions (that cannot have an anticipatory function or be shifted to attain one), and we additionally have case inflections attached to phrase constituents that do result in a relatively early recognition of structure, and thus meaning.

5 We would make this argument even if it will, in most cases, be clear from the context what relation is expressed.

6 As Schützler (2018) shows, the postpositional and prepositional use of *notwithstanding* in English correlates with the length of the NP complement in a way that is in accordance with the CP. Nevertheless, no such positional variants exist in the Estonian constructions we investigated.

The notion of explicitness is central to the theory developed by Hawkins (1990, 1992, 1994, 2004), where it is mainly about the early recognition of syntactic structures and may have less to do with the precise form of a marker. Our results can, therefore, be explained by Hawkin's (2004) view of efficiency – a concept more difficult to define than complexity. According to Hawkins (2004: 9), efficiency is increased by using linguistic forms "so as to provide the earliest possible access to as much of the ultimate syntactic and semantic representation as possible". This means that Estonian locative case affixes maximize the on-line property assignment better than Estonian postpositions. In line with Hawkins (2004), we propose that with long and complex Landmark phrases, the case construction is the (most) efficient structure since it provides the earliest possible access to properties in the proposition to be communicated, i.e. expressing exterior location in our case. Both of Hawkins' (2004: 25) slogans – "Express the most with the least" and "Express it earliest" apply. This ties in with the two forces at play in language mentioned by Dahl (2004: 135): effectiveness or clarity and efficiency or economy. The first force implies that "speakers seek to make themselves understood and therefore strive for maximally effective messages"; the second is that "there is a general tendency not to expend more energy than is strictly necessary and therefore to prefer economical forms to more elaborate ones".

In our attempt to focus our notion of complexity on the properties of Estonian, we also turned to the Dependency Locality Theory (e.g. Gibson 2000; Temperley 2007; Futrell et al. 2020) as well as Hawkins's Minimize Domains Principle (Lohse et al. 2004) but did not find them helpful in our particular case, since they too largely depend on word-order variation. For Estonian data, the jury is still out: We lack psycholinguistic evidence to either confirm or refute the applicability of the claims put forward by morphological and syntactic processing theories for a language typologically different from English. This makes an explicit comparison of performance theories complex: What the different processing theories tell us about English and other Indo-European languages may not apply to non-Indo-European linguistic structures. Current psycholinguistic processing models are based on structural orderings, i.e. alternative orderings of the same constituent – something the Estonian alternations are not.

In sum, our study has provided evidence that the CP, as formulated by Rohdenburg, does not generalise sufficiently to other languages to warrant the status of a 'principle'. In particular, it may fail to replicate in contexts where morphological weight and early placement do *not* coincide and/or in highly inflectional languages. Future work in the field of Cognitive Linguistics will hopefully shed more light on these and related issues.

Data availability statement

The dataset used in the present study has been published as *Background Data for: The complexity principle and the morphosyntactic alternation between case affixes and postpositions in Estonian* at the TROLLing repository (Klavan 2023, https://doi.org/10.18710/KDSZEP). R-scripts used in the analysis and visualization of the data can be retrieved from https://osf.io/sw73b/, including the graphics files. Note that the format of variables in the original dataset is not the same as the one used in this study: This article depends on the original (published) dataset in combination with the analytic scripts, which involve some re-formatting of variables. For the entire repository, we selected a CC BY 4.0 licence (https://creativecommons.org/licenses/by/4.0/) and marked the respective figures in the captions thus:  .

Appendix

A. Functions of the Estonian exterior cases (adapted from Erelt et al. 2007: 249–251)

ALLATIVE FUNCTION	EXAMPLE SENTENCE	POSTPOSITIONAL ALTERNATIVE	ENGLISH TRANSLATION
Direction of location	Mari pani vaasi **lauale.**	Mari pani vaasi **laua peale.**	'Mari put the vase on(to) the table.'
Time	Koosolek viidi üle **neljapäevale.**	Koosolek viidi üle **neljapäeva peale**.	'The meeting has been moved to Thursday.'
State	Tüdruku nägu läks **naerule**.	not attested	'The girl started to laugh.'
Addressee	Mari rääkis **Jürile** kõik ära.	not attested	'Mari told Jüri everything.'
Experiencer	**Mulle** meeldib siin elada.	not attested	'I like living here.'
Object of action	Ta lootis **sõpradele**.	Ta lootis sõprade peale.	'He counted on friends.'
Object of emotions	Mihkel on **sõbrale** kade.	Mihkel on **sõbra peale** kade.	'Mihkel is jealous of his friend.'
Without clear meaning	Järgenege mulle.	not attested	'Follow me.'
ADESSIVE FUNCTION	**EXAMPLE SENTENCE**	**POSTPOSITIONAL ALTERNATIVE**	**ENGLISH TRANSLATION**
Location	Vaas on **laual**.	Vaas on **laua peal**.	'The vase is on the table.'
Time	Nad sõidavad **neljapäeval** maale.	not attested	'They are driving to the country on Thursday.'

(continued)

ALLATIVE FUNCTION	EXAMPLE SENTENCE	POSTPOSITIONAL ALTERNATIVE	ENGLISH TRANSLATION
State	Jüri vaatas meid **naerul** näoga.	not attested	'Jüri looked at us with a laughing face.'
Possessor	**Maril** on kaks last.	not attested	'Mari has two children.' (lit. 'On Mary are two children.')
Agent with finite verb forms	See asi ununes **mul** kiiresti.	not attested	'I quickly forgot about that thing.'
Instrument	Mari mängib **klaveril** mõnd lugu.	Mari mängib **klaveri peal** mõnd lugu.	'Mari is playing some tunes on the piano.'
Manner	Mari kuulas kikkis **kõrvul**.	not attested	'Mari listened with her ears pricked up.'
ABLATIVE FUNCTION	**EXAMPLE SENTENCE**	**POSTPOSITIONAL ALTERNATIVE**	**ENGLISH TRANSLATION**
Source of location	Mari võttis vaasi **laualt**.	Mari võttis vaasi **laua pealt**.	'Mari took the vase off the table.'
Source	Mari kuulis seda **Jürilt**.	not attested	'Mari heard it from Jüri.'
Modifier of a noun	**Elukutselt** on ta insener.	not attested	'He is an engineer by profession.'

B. Basic table of model coefficients

More detailed summaries are shown in the online materials. The levels of predictors are as follows (cf. Table 1):

POSITION: reference = LM follows TR; predicted = LM precedes TR
COMPLEXITY: reference = simple; predicted = compound
LENGTH: continuous, centred ($\log_2$ syllables)
FREQUENCY: continuous, centred ($\log_2$ pmw)
FUNCTION: reference = adverbial; predicted = modifier

	Construction		
	Lative	**Locative**	**Separative**
Fixed part			
INTERCEPT	−0.24	1.22	0.19
COMPLEXITY	−0.80	−1.88	−1.98

(continued)

	Construction		
	Lative	**Locative**	**Separative**
FUNCTION	−1.18	−1.23	−0.48
LENGTH	−0.95	−1.71	−0.39
FREQUENCY	−0.11	−0.33	−0.24
POSITION	0.01	0.23	0.39
n	1,000	1,000	1,000
Random part			
LM_LEMMA: SD	1.37	1.43	1.92
LM_LEMMA: *n*	592	438	528
VERB_LEMMA: SD	1.53	0.61	1.48
VERB_LEMMA: *n*	224	279	252
Priors (same for all alternations)			
intercept		normal(0, 1.5)	
b		normal(0, 1.5)	
SD		normal(0, 1.0)	

References

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

Aristar, Anthony Rodrigues. 1997. Marking and hierarchy types and the grammaticalization of case-markers. *Studies in Language* 21(2). 313–368.

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

Barr, Dale J., Roger Levy, Christoph Scheepers & Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68(3). 255–278.

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.

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.

Bürkner, Paul-Christian. 2020. *brms. Bayesian Regression Models using 'Stan'*. R-package version 2.12.0. Available at: https://cran.r-project.org/web/packages/brms/index.html.

Bybee, Joan L. 2013. Usage-based theory and exemplar representations of constructions. In Thomas Hoffmann & Graeme Trousdale (eds.), *The Oxford handbook of construction grammar*, 49–69. Oxford: Oxford University Press.

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

Dahl, Östen. 2004. *The growth and maintenance of linguistic complexity*. Amsterdam: John Benjamins.

Erelt, Mati. 2003. Syntax. In Mati Erelt (ed.), *Estonian language*, 93–129. Tallinn: Estonian Academy Publishers.

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

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

Futrell, Richard, Roger P. Levy & Edward Gibson. 2020. Dependency locality as an explanatory principle for word order. *Language* 96(2). 371–412.

Gibson, Edward. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In Alec Marantz, Yasushi Miyashita & Wayne O'Neil (eds.), *Image, language, brain: Papers from the first mind articulation project symposium*, 95–126. Cambridge, MA: The MIT Press.

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.

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

Haspelmath, Martin. 2006. Against markedness (and what to replace it with). *Journal of Linguistics* 42(1). 25–70.

Hawkins, John A. 1990. A parsing theory of word order universals. *Linguistic Inquiry* 21(2). 223–261.

Hawkins, John A. 1992. Syntactic weight versus information structure in word order variation. *Linguistische Berichte* 4(Special issue). 196–219.

Hawkins, John A. 1994. *A performance theory of order and constituency*. Cambridge: Cambridge University Press.

Hawkins, John A. 2004. *Efficiency and complexity in grammars*. Oxford: Oxford University Press.

Heinze, Georg, Christine Wallisch & Daniela Dunkler. 2018. Variable selection – a review and recommendations for the practicing statistician. *Biometrical Journal* 60(3). 431–449.

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.

Hinrichs, Lars & Benedikt Szmrecsanyi. 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.

Kallas, Jelena & Kristina Koppel. 2018. *Eesti keele ühendkorpus 2017*. [Estonian National Corpus 2017]. Tallinn: Center of Estonian Language Resources.

Klavan, Jane. 2012. *Evidence in linguistics: Corpus-linguistic and experimental methods for studying grammatical synonymy*. Tartu: University of Tartu Press Dissertationes Linguisticae Universitatis Tartuensis 15.

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.

Klavan, Jane. 2023. Background Data for: The complexity principle and the morphosyntactic alternation between case affixes and postpositions in Estonian. https://doi.org/10.18710/KDSZEP, DataverseNO, V1.

Klavan, Jane, Maarja-Liisa Pilvik & Kristel Uiboaed. 2015. The use of multivariate statistical classification models for predicting constructional choice in spoken, nonstandard 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 soomeugri keeleteaduse ajakiri. Journal of Estonian and Finno-Ugric Linguistics* 8(2). 59–91.

Klavan, Jane, Tanel Alumäe & Arvi Tavast. 2020. Eesti keele väliskohakäänete kasutus poolspontaanses kõnes automaatse transkriptsiooni põhjal. [Analysis of Estonian external locative cases in semi-spontaneous speech using an automatic transcription system]. *Keel ja Kirjandus* 63(8–9). 757–774.

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.

Lestrade, Sander. 2010. *The space of case*. Nijmegen: Radboud University Dissertation.

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]. Tartu: University of Tartu Press Dissertationes Filologiae Estonicae Universitatis Tartuensis 16.

Lohse, Barbara, John A. Hawkins & Thomas Wasow. 2004. Domain minimization in English verb-particle constructions. *Language* 80(2). 238–261.

Mithun, Marianne. 1987. Is basic word order universal? In Russell Tomlin (ed.), *Coherence and grounding in discourse*, 281–328. Amsterdam: John Benjamins.

Mondorf, Britta. 2009. *More support for more-support: The role of processing constraints on the choice between synthetic and analytic comparative forms*. Amsterdam: John Benjamins.

R Core Team. 2019. *R: a language and environment for statistical computing*. Version 3.6.2. Available at: https://www.r-project.org/.

Rohdenburg, Günter. 1996. Cognitive complexity and increased grammatical explicitness in English. *Cognitive Linguistics* 7(2). 149–182.

Rohdenburg, Günter. 2002. Processing complexity and the variable use of prepositions in English. In Hubert Cuyckens & Günter Radden (eds.), *Perspectives on prepositions*, 79–100. Tübingen: Niemeyer.

Rohdenburg, Günter. 2021. The Complexity Principle at work with rival prepositions. *English Language and Linguistics* 24(4). 769–800.

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

Rosenbach, Anette. 2014. English genitive variation–the state of the art. *English Language and Linguistics* 18(2). 215–262.

RStudio Team. 2009–2019. RStudio: Integrated Development for R. Version 1.2.5033. Available at: http://www.rstudio.com/.

Sarkar, Deepayan. 2018. *lattice: Trellis Graphics for R*. R-package version 0.20-38. Available at: https://cran.r-project.org/web/packages/lattice/index.html.

Schützler, Ole. 2018. Grammaticalization and information structure: Two perspectives on diachronic changes of *notwithstanding* in written American English. *English Language and Linguistics* 22(1). 101–122.

Sönning, Lukas. Forthcoming. Drawing on principles of perception: The line plot. In Lukas Sönning & Ole Schützler (eds.), *Data visualization in corpus linguistics: Critical reflections and future directions*. Helsinki: VARIENG.

Stan Development Team. 2019. *Stan modeling language users guide and reference manual*. version 2.26. Available at: http://mc-stan.org.

Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Talmy, Leonard. 1983. How languages structure space. In Herbert L. PickJr. & Linda P. Acredolo (eds.), *Spatial orientation: Theory, research and application*, 225–282. New York: Plenum Press.

Talmy, Leonard. 2000. *Toward a cognitive semantics*, vol. 2. Cambridge, MA: The MIT Press.
Temperley, David. 2007. Minimization of dependency length in written English. *Cognition* 105(2). 300–333.
Tizón-Couto, David & David Lorenz. 2021. Variables are valuable: Making a case for deductive modelling. *Linguistics* 59(5). 1279–1309.
Veismann, Ann & Mati Erelt. 2017. Kaassõnafraas. [Adpositional phrase. In Mati Erelt & Helle Metslang (eds.), *Eesti keele süntaks* [Estonian Syntax], 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.
Wasow, Thomas. 1997. Remarks on grammatical weight. *Language Variation and Change* 9(1). 81–105.
Wasow, Thomas. 2002. *Postverbal behaviour*. Stanford, CA: CSLI Publications.
Wolk, Christoph, Joan Bresnan, Anette Rosenbach & Benedikt Szmrecsanyi. 2013. Dative and genitive variability in Late Modern English. *Diachronica* 30(3). 382–419.