Nor Jnl Ling 40.3, 313–349 © Nordic Association of Linguists 2017 doi:10.1017/S0332586517000191

Lindström, Liina & Kristel Uiboaed. 2017. Syntactic variation in 'need'-constructions in Estonian dialects. *Nordic Journal of Linguistics* 40(3), 313–349.

Syntactic variation in 'need'-constructions in Estonian dialects

Liina Lindström & Kristel Uiboaed

The article contributes new data and findings to the growing field of corpus-based dialect syntax research. The focus of the paper is on variation in 'need'-constructions (*tarvis*/*vaja olema* + nominal complement/infinitive 'need to') based on the corpus of Estonian dialects. Our purpose was to demonstrate the complex nature of syntactic variation, constrained geographically, individually or by language-internal factors. The study takes a corpus-based quantitative approach to observing the geographical spread of linguistic units. We apply conditional inference tree and random forests models to capture the (co)varying parts of the construction studied. Our results show that variation in different parts of constructions is influenced by different factors, both geographical and language-internal. Lexical variation (adverb *tarvis* 'need' or *vaja* 'need') and omission of the copula are clearly geographically distributed, while omission of the experiencer is determined mainly by language-internal factors. However, the study has also found extensive inter-individual differences.

Keywords: dialect syntax, Estonian dialects, language contact, modal constructions, syntactic variation

Liina Lindström, University of Tartu, Institute of Estonian and General Linguistics, Jakobi 2, 51014 Tartu, Estonia. liina.lindstrom@ut.ee
Kristel Uiboaed, University of Tartu, Institute of Estonian and General Linguistics, Jakobi 2, 51014 Tartu, Estonia. kristel.uiboaed@ut.ee

1. INTRODUCTION

Over the past two decades the focus of dialect research has shifted from phonology, lexicology and morphology to other domains of grammar, such as syntax. Corpus-based research on dialect syntax has become a particularly lively subject area where new resources and methodology are actively applied and developed (e.g. Grieve 2009, Szmrecsanyi 2013, Wolk & Szmrecsanyi 2016). New resources, i.e. large dialect corpora, enable the application of methods widely used in statistics and corpus linguistics – more generally in the usage-based framework – to complex and heterogeneous dialect data. The analysis of this kind of data presupposes rigorous statistical methodology in order to capture new information about the nature of variation, observing the effect of local language contacts and the emergence of linguistic constructions more generally.

https://doi.org/10.1017/S0332586517000191 Published online by Cambridge University Press

This study continues a similar line of research and utilizes corpus data to examine syntactic variation in Estonian dialects by considering the frequency of syntactic phenomena and applying quantitative methods. We analyse Estonian 'need'-constructions, using the data drawn from the Corpus of Estonian Dialects. The main purpose of this study is to provide insight into geographic variation in syntax and its complexity, and to detect the potential sources of the variation. We demonstrate that when analysing the construction that varies, we have to consider, along with geography, language-internal syntactic and semantic factors, including construction-internal features, since some items within the construction affect variation in other parts of the same construction. Our focus, however, is on dialectal distribution; in other words, our main purpose is to determine what part of the variation can be explained by inter-dialectal differences.

By including frequencies in the research paradigm, we gain a substantial amount of new information about spatial variation in syntactic phenomena. Frequency-inferred tendencies can often be explained by language contact, and, thus, are far from coincidental. Usage patterns already existing in a language (or dialect) can acquire higher usage frequency due to similar patterns in a contact (or model) language that the speakers are frequently confronted with (Heine & Kuteva 2005:47). Even if some phenomenon exists in various areas, its usage frequency can be substantially lower outside the core of the observed contact region (see Koptjevskaja-Tamm & Wälchli 2001:627). Change in the usage frequency of some (previously existing) pattern has been considered the typical example of contact-induced grammatical transfer (Heine & Kuteva 2005:48). The Estonian 'need'-constructions under investigation share many properties with Russian and Latvian (see Section 3), and we expect that relative frequencies may reflect contacts with the respective neighboring languages.

We look closely at the variation in 'need'-constructions in Estonian dialects containing the predicates *tarvis*/*vaja* (*olema*) 'need', as in examples (1) and (2) (for abbreviations see the list at the end of this paper, immediately before References).

(1) Mu-l on vaja kooli minna.
I-ADE *be.3*SG *need* *school.*ILL *go.*INF
'I need to go to school.'

(2) Mu-l on tarvis uut arvuti-t.
I-ADE *be.3*SG *need* *new.*PRT *computer*-PRT
'I need a new computer.'

On the basis of form–meaning pairings, we have distinguished two basic 'need'-constructions in the sense of Construction Grammar (e.g. Croft 2001). The infinitival construction (1) takes an infinitival complement e.g. *kooli minna* 'go to school' and has a modal meaning, marking necessity or obligation. It is clearly grammaticalized to the modal domain: the modal adverb *tarvis*/*vaja*, copula *ole-* 'be' and the infinitival

https://doi.org/10.1017/S0332586517000191 Published online by Cambridge University Press

complement together act as a modal predicate. The nominal complement construction (2) expresses a need for or lack of something and is less grammaticalized in this respect; it can therefore also be called a pre-modal construction. (See Lindström & Vihman 2017 for more information about the same constructions in standard Estonian.) These two basic constructions are the most common in the data.

However, there is more variation within the constructions, and it is not always easy to decide whether we are dealing with intraconstructional variation or with a different construction, because all the parts of the constructions may vary without obvious differences in meaning. In both basic constructions, the semantic core is a modal adverb, either *tarvis* or *vaja* 'need'. They have been considered synonymous; no previous research has highlighted possible semantic and/or dialectal differences between the use of *tarvis* and *vaja*. The experiencer argument in the constructions may occur in the adessive, as *mul* 'I' in (1) and (2), but the experiencer may also be case-marked with the allative, as in (3), or may be omitted (marked by *0*), as in (4).

(3) kui su-lle ädäste tarviss omm siss mine ja
if you-ALL so_much need be.3SG then go.IMP and
võtta tu raha (Tartu)[1]
take.IMP this money
'if you need (it) so much, go and take this money'

(4) ommukku leit-si-n et 0 vaea riide-sse aea-da
morning.GEN find-PST-1SG that need cloth-ILL drive-INF
ja minna minekki-t. (Eastern)
and go.INF going-PRT
'In the morning I found that (I) needed to get dressed and get going.'

The constructions include a copula *ole-* 'be' in the singular third person (3), but the verbs *minema* 'go' or *tulema* 'come' may act similarly to the copula. The copula may also be omitted, as is seen in (4). The complement of the construction can be clausal or may even be omitted, in addition to the basic patterns (infinitival and nominal, as in (1) and (2)).

We focus on variation that can be analysed statistically: lexical variation (using *tarvis* or *vaja*), variation in the copula (especially ellipsis of the copula), and variation in experiencer-marking (especially ellipsis of the experiencer).

Previous variationist research has emphasized the importance of individual internal variation (e.g. Van de Velde & van Hout 1998, Tagliamonte & Baayen 2012). To capture more fine-grained information about variation, we conducted all our analyses in two ways – with and without the individual speaker variable. Our hypothesis was that the construction in question has a distinct geographic distribution, and that individual differences are secondary.

The paper is organized as follows: In Section 2, we provide a brief overview of dialect syntax approaches, especially with regard to the data sources that have been used in the field, and a brief overview of work in Estonian dialect syntax to date. In Section 3, we present an overview of Estonian dialects, language contacts, and the ways in which 'need'-constructions are used in contact languages. Section 4 describes the data and methods applied and Section 5 presents the results of the study.

2. DIALECT SYNTAX

During the past two decades, both the generativist paradigm and linguistic typology have found in dialects a useful source of new information about linguistic varieties (Kortmann 2010, Bucheli Berger, Glaser & Seiler 2012). One of the reasons why dialect syntax was a rather neglected field before the 1990s was probably the nature of syntactic variation, which – compared to phonological and lexical variation – is much subtler and less salient; it is also in many cases less categorical and rather a matter of frequency (Kortmann 2010). Syntactic phenomena can be better described in terms of preferences and preference patterns, rather than the simple presence or absence of certain features (Kortmann 2010), thus syntactic variation is often probabilistic rather than categorical in nature (Wolk et al. 2013).

Methodologically, research on dialect syntax has developed in two main directions in terms of the sources of data used: studies using linguistic questionnaires and those based on dialect corpora. We will next briefly discuss these two methods.

The use of QUESTIONNAIRES as a method for dialect syntax research has been widespread in studying, for example, Dutch, German and the Nordic languages (Vangsnes 2007, Bucheli Berger et al. 2012) but has been less common in the case of many other languages, including Estonian and Finnish. Traditionally, questionnaire data have been presented in the form of linguistic atlases, maps, etc. This method allows us to take into account rarely occurring syntactic phenomena, and to ask respondents about, for example, semantic differences, individual preferences, and different contexts. This method, however, needs adequate databases, either from some earlier period or new data. The latter presupposes the existence of vital dialect-speaking communities, who can provide reliable data. Nowadays, this is no doubt problematic for many languages, because of leveling processes affecting traditional dialects. Estonian dialects, the focus of the current paper, have undergone a rapid leveling process of traditional geographic dialects, and systematic syntactic questionnaire data from earlier periods are lacking. We can only find a few syntactic features form the Estonian Dialect Atlas, compiled by Andrus Saareste. These atlases (Saareste 1938, 1955) contain mostly lexical data.

Recently, DIALECT CORPORA have become available for many languages, which enables the study of syntactic variation in its natural context in a quantitative manner. Such a corpus has also been compiled for Estonian dialects, as we have plenty of recorded dialect interviews since the 1950s (see Section 4.1 below). Compared to questionnaires, corpus data have both advantages and disadvantages. The apparent

https://doi.org/10.1017/S0332586517000191 Published online by Cambridge University Press

advantage of a dialect corpus is that it affords the possibility of collecting frequency data, making statistical analysis much easier. Szmerczsanyi (2013:4) has pointed out that frequency is a better indicator of variational patterns than the traditionally used dialect atlases, where only discrete classifications are used. Using frequency data, we can more adequately explain the typical and atypical features of a particular dialect, which is especially important in areas where two or more competing patterns are attested. The apparent disadvantage of dialect corpora is the limited amount of data, especially when dealing with spoken vernaculars and texts from many different (sub-)dialects. Dialect corpora thus usually enable the analysis of frequently occurring linguistic phenomena, but are seldom useful in studying rare and infrequent phenomena. This is a problem we are faced with in the present study as well: while the total extent of annotated texts was approximately 900,000 words, we still do not have enough data from each region to study less frequently occurring variants of the construction. We also have to keep in mind that corpora never give every possible context or usage of a linguistic phenomenon; we have to work with the data we are able to retrieve.

In Estonian dialectology, syntactic phenomena have typically been outside the scope of research. There has been some work on Estonian dialect syntax, but most of it has concerned only a particular syntactic phenomenon in a particular (sub-)dialect (e.g. Koit 1963; Neetar 1964, 1970; Lindström et al. 2009; Mets 2010; Uiboaed 2010; Kehayov, Lindström & Niit 2011; Velsker 2013, Metslang & Lindström 2017). Some syntactic features are also listed in descriptions of dialects (e.g. Must 1987, Must & Univere 2002, Pajusalu et al. 2009). More exhaustive studies are available only on agreement phenomena (Nurkse 1937, Neetar 1964) and evidential constructions (Kask 1984). The first attempt to apply the principles of aggregate frequency-based dialectology to Estonian verbal constructions was conducted by Uiboaed (Uiboaed 2013, Uiboaed et al. 2013). More recently, the corpus-based approach to syntactic variation in Estonian dialects has been applied to e.g. the frequency distribution of perfect and pluperfect (Lindström et al. 2015, Lindström et al. 2017), variation in locative constructions marking 'on' (Klavan, Pilvik & Uiboaed 2015), use of ambipositions (Ruutma et al. 2016), constructions with action nominal (Pilvik 2016, 2017), and use of partitive subjects (Lindström 2017).

3. ESTONIAN DIALECTS, LANGUAGE CONTACTS AND 'NEED'-CONSTRUCTIONS IN CONTACT LANGUAGES

Estonian is a Finno-Ugric language belonging to the Finnic branch. Its closest relatives are Livonian and Votic, which at present are nearly extinct. The closest languages to Estonian still used for everyday communication are Finnish, Karelian, and Veps. Estonian has a complex morphological system, which is typical of the Finno-Ugric languages (Erelt, Erelt & Ross 2000).

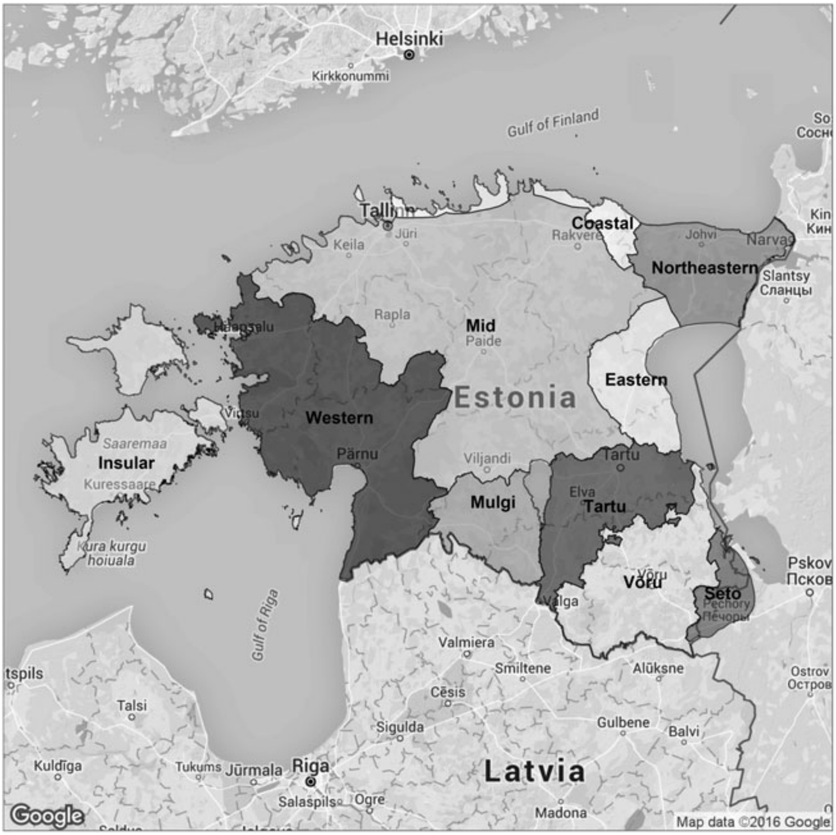

Figure 1. Estonian dialect areas.

The area where Estonian is spoken is relatively small, but differences among the traditional dialects are substantial. There are slightly different classifications of Estonian dialects available. The classification that is used in the corpus of Estonian dialects establishes three main dialect groups: the North Estonian group, which includes the Insular, Western, Mid, and Eastern dialects; the South Estonian group, consisting of the Mulgi, Tartu, Seto, and Võru dialects; and the Northeastern-Coastal dialect group, which includes the Coastal and Northeastern dialects. These three groups can be divided into more than 100 sub-dialects. The map in Figure 1 shows the main Estonian dialect areas. Traditional dialect classifications distinguish most significantly between northern and southern dialects; the greatest differences are in phonology, morphology and lexis. South-Estonian dialects are sometimes treated as a separate language.

Estonian has been heavily influenced by the Indo-European languages, and has acquired a number of features that are characteristic of Standard Average European

https://doi.org/10.1017/S0332586517000191 Published online by Cambridge University Press

(Metslang 2009). Estonian has also had long-lasting contacts with languages around the Baltic Sea, forming the Circum-Baltic language area (see Dahl & Koptjevskaja-Tamm 2001, Wälchli 2011). The Circum-Baltic area includes Finnic languages, Russian, Baltic languages (Latvian and Lithuanian), Germanic languages (particularly Swedish, German, Low German) and several other languages spoken in the region (Koptjevskaja-Tamm & Wälchli 2001, Wälchli 2011). The Circum-Baltic language area has been described as a complex linguistic community where mutual influences of individual languages are constant and there is no single dominating language; therefore, it is also referred to as a contact superposition zone rather than *Sprachbund* (Koptjevskaja-Tamm & Wälchli 2001:626; Wälchli 2011:325). Estonian belongs to the eastern part of the Circum-Baltic languages together with Baltic, Finnic, and Slavic languages (especially Russian). Although these languages belong to different language families, many morphosyntactic similarities have been attested among them (see e.g. Dahl & Koptjevskaja-Tamm 2001; Vaba 2011; Seržant 2012, 2015a, b; Klaas-Lang & Norvik 2014).

The 'need'-constructions discussed in this article have some obvious semantic parallels in Baltic and Slavic languages. The Latvian counterpart mainly marks necessity, both dynamic and deontic, with a preference for deontic use (see Holvoet 2001:28–32, 2009:209), similarly to Estonian; and the same appears in Russian (Wade 2011:313). In all three languages, the modal 'need'-predicates can also be used with a nominal complement (pre-modal usage), expressing need for something or lack of something.

In addition, there are clear structural parallels between Estonian, Latvian and Russian 'need'-constructions. Firstly, the use of modal adverbs and modal adjectives in these constructions is characteristic to Russian and other East Slavonic languages (e.g. *nado*, *nuzhno* 'need' in Russian), while West Slavic languages prefer modal verbs (Hansen 2005). In Balto-Finnic languages, both modal adverbs and modal verbs are used, but modal adverbs are widely used especially in Estonian as well as some eastern Balto-Finnic languages that have been under the strong influence of Russian (Kehayov & Torn-Leesik 2009). Another parallel between Estonian, Latvian and Russian is related to the marking of the experiencer: in Russian and Latvian 'need'-constructions, dative case is used to mark the modal experiencer, while in Estonian and eastern Balto-Finnic languages, adessive and allative are widely used as a counterpart of the dative in experiential constructions (see also Seržant 2015b). In all these languages, the dative(-like) experiencer can also be easily omitted in modal constructions (Kehayov & Torn-Leesik 2009, Hansen 2014). A third parallel can be found comparing Estonian and Latvian: both use the same stem *vaja*. In Estonian, *vaja* is used as a modal adverb, while Latvian uses the verb *vajadzēt* for marking necessity. The stem *vaja-* has been borrowed from Livonian to Latvian (Kalnača 2013), and later verbalized (Seržant & Bjarnadóttir 2014). The Latvian verb *vajadzēt* can only be used in third person singular form, while the Estonian construction uses

https://doi.org/10.1017/S0332586517000191 Published online by Cambridge University Press

the copula *olema* 'be' in third person singular form. Russian also uses the copula in third person form, when it is present in the clause. The parallels between the languages are exemplified in (5)–(7).

(5) Mu-l on vaja tee-d juua. (Estonian)
I-ADE be.3SG need tea-PRT drink.INF
'I need to drink (some) tea.'

(6) Man vajag dzert tēju. (Latvian)
I:DAT need:PRS.3SG drink:INF tea:F:ACC
'I need to drink (some) tea.'

(7) Mn'e nado/nuzhno po-pit' čaj. (Russian)
I:DAT need drink:INF tea:M:ACC
'I need to drink (some) tea.'

Some of these features can also be found in Finnish necessive constructions, but to a lesser degree. Finnish mostly uses modal verbs (such as *täytyy*, *pitää*, *tarvita* 'need, have to'), and fewer modal adverbs than Estonian (e.g. *pakko* means mainly obligation). The Finnish modal subject has non-nominative marking, but unlike Estonian, mostly genitive is used (in dialects and colloquial speech, adessive can also be used, see Laitinen 1992:112–113). In Finnish, ellipsis of the experiencer argument is common, and the elliptical NP refers to people familiar from the context, either the discourse participants or the narrative protagonist, or it remains generic (see Laitinen 1992:109–110). Experiencer ellipsis in necessive constructions referring to discourse participants has also been noted in the Finnish comprehensive grammar (ISK:1291).

In most of the studies on language contact, Estonian is treated as a whole and the language-internal variation is not accounted for. However, a multitude of phenomena in Estonian display significant regional differences which can be explained by contacts that have taken place locally, i.e. they have affected only certain areas. Our previous findings (Lindström et al. 2015, Lindström et al. 2017) have shown that dialects may differ remarkably by usage frequency of relatively old, contact-induced phenomena common to all Estonian dialects: usage frequency of perfect and pluperfect is notably higher than average in dialects that have had long-lasting contacts with languages where the corresponding construction exists, and notably lower in areas with more contacts with languages where the corresponding construction is missing.

We expect that some of the variation we detect in 'need'-constructions is due to (or reinforced by) ongoing language contact which have occurred more locally. We assume that features that are usual also in Russian and/or Latvian, are more frequent in eastern and southern part of Estonia, respectively. Language contact thus serves as an explanatory factor for construction-internal variation, and it may explain differences between the dialects. The explanatory power of language contact, however, is hard to operationalize for quantitative analysis because the relevant contact period, nature and intensity of contacts are hard to determine and may vary largely. We can estimate the role of contacts only indirectly, taking into account geographical distance and earlier information about contacts.

However, as we will show in the following sections, language contacts do not explain all the variation; construction-internal factors also play a role, as well as individual differences.

4. DATA AND METHODS

4.1 Corpus of Estonian Dialects

Our data come from the Corpus of Estonian Dialects (CED; http://www.murre.ut.ee/estonian-dialect-corpus/), compiled at the University of Tartu in collaboration with the Estonian Language Institute. The CED represents spoken dialects recorded during the years 1957–1980, and includes dialect interviews from all Estonian dialect areas. Since the recordings selected for the CED are the oldest tape-recorded dialect texts available, the corpus represents relatively old dialect speech. More recently Estonian dialects have quickly leveled, and most of them are no longer in active use. The informants have been selected on the basis of classic criteria: they are local people, typically not highly educated, and have not moved around during their lives. The texts in the corpus represent dialect interviews, which often include long monologue passages. The topics of the interviews typically deal with such topics as the respondent's personal life, family, lifestyle and working methods in the past, or episodes in their lives.

The CED consists of five parts: (i) dialect recordings, (ii) phonetically transcribed dialect texts (in Finno-Ugric transcription system), (iii) dialect texts in simplified transcription, (iv) morphologically annotated texts in XML-format, and (v) a separate database containing information about informants and recordings. The morphologically annotated part of the corpus contains 940,000 tokens.

4.2 Data

The data for *tarvis* and *vaja olema* constructions are extracted from the morphologically annotated part of the CED. The final dataset comprises 355 instances of modal predicates containing either *tarvis* or *vaja*, from 135 individuals (16 are variable individuals). The final dataset contains all sentences retrieved and possible to code for the chosen factors, so we are dealing with a holistic sample. The next sections describe the factors coded and their levels for our analysis.

https://doi.org/10.1017/S0332586517000191 Published online by Cambridge University Press

4.2.1 Modal adverb

Constructions are extracted from the CED by the lexical item used in the clause. The modal adverb is the only obligatory part of the construction, and is the binary dependent variable (levels *tarvis* and *vaja*) in our first analysis. In the Estonian reference grammar (Erelt et al. 1993) the modal predicates *tarvis* and *vaja olema* are listed as synonymous predicates, expressing the same meanings and functions.

4.2.2 Dialect and individual

The basic geographical unit in this study is the dialect, as represented in Figure 1 above. The same division of dialects is used in the CED. For practical reasons we could not use smaller units (sub-dialects), as the number of usages of 'need'-constructions distributed over the smaller areas is too small for quantitative study; we have therefore aggregated over the sub-dialect areas. Hence the predictor *dialect* has ten levels (Eastern, Coastal, Insular, Mid, Mulgi, Northeastern, Seto, Tartu, Võru, Western). In addition, all the utterances contain the code for the informant who produced the sentence (predictor *individual*). In order to explain potential individual preferences, informants are traditionally included in predictors set in variationist and corpus-based studies (e.g. Guy 1980, Van de Velde & van Hout 1998, Tagliamonte & Baayen 2012). Although, including other sociolinguistic variables would follow the traditional line of research and likely contribute some interesting information, we have omitted these in the present analysis for technical reasons: the dataset is fairly small, heterogeneous and contains cells with missing information. All this would introduce additional problems in statistical analysis.

4.2.3 Copula verb

The most commonly used verb in 'need'-constructions is the copula *olema* 'be' (as in examples (1) and (2) above). In addition, the verbs *tulema* 'come' and *minema* 'go' can also carry a copular function in the construction, but are rather infrequent. (See examples (7) and (8) in Section 5.2 below.) The copula may also be omitted. The predictor *mainverb* thus has four levels in our data, according to the lexical verb used (*olema*, *minema*, *tulema*) or ellipsis (coded as *0*). The omission of the copula is common in Russian (Kehayov 2009), and this is a reason why we expect frequent copula ellipsis in eastern dialects that have had more contacts with Russian.

4.2.4 Marking of the experiencer

The experiencer argument is case-marked either by the adessive (*ade*) or by the allative (*all*), but can also be left unexpressed (*0*); hence the predictor *Experiencer* consists of these three levels. To explore the factors that potentially influence the omission of the experiencer argument, we have created the variable *exp_is*, which

https://doi.org/10.1017/S0332586517000191 Published online by Cambridge University Press

shows the presence of the experiencer irrespective of its case-marking. This is thus a binary variable, with two levels (*yes* and *no*). Since the omission of the experiencer in modal constructions is common in Slavic and Baltic languages (Hansen 2014), we expect that the rate of experiencer omission is higher in eastern and southern dialects.

4.2.5 Complement

The 'need'-construction can take a nominal complement (nom), marked mostly with the partitive case, as in example (2) above; but nominatives and infinitival complements (*inf*), as in (1) above, also occur. The complement may also be elided (*0*), in cases where the object needed or the action required is clear from the discourse or from the more general context. The infinitival complement occurs in modal constructions marking either dynamic or deontic necessity. Thus clauses with the infinitival complement represent more grammaticalized usages than clauses taking the nominal complement.

4.2.6 Polarity

All clauses were marked for polarity, with binary predictor levels: *aff* (affirmative polarity) and *neg* (negative polarity). In *tarvis*/*vaja* constructions, the negative marker can be attached only to the copula; thus the predictor could be relevant in the analysis of the occurrence of the copula verb. We expect that negative polarity increases the probability of explicit use of copula.

4.2.7 Tense

Tense markers can be attached only to the copula, and the predictor is thus expected to be relevant in the analysis of the occurrence of the copula verb. We have distinguished between present tense (*pr*) and past tenses (*ipf*). The latter includes all past tenses, i.e. simple past, perfect and pluperfect. We expect that past tense (which is marked on the copula) increases the probability of explicit use of copula.

In Section 5, we take a closer look at the choice between the modal adverbs *tarvis* and *vaja*, variation in the use of the copula (focusing on ellipsis of the copula), and variation in the marking of the experiencer argument. Our primary objective is to investigate the main predictors that best predict the choice between alternative constructions. The principal question is, whether the predictor dialect is statistically significant when it comes to predicting the construction usage, i.e. whether the chief constraints on usage are geographical or whether other (possibly) significant language-internal predictors can be identified. We use the term DEPENDENT VARIABLE for the phenomenon under investigation and the term PREDICTOR and PREDICTOR LEVELS for explanatory variables and their levels.

https://doi.org/10.1017/S0332586517000191 Published online by Cambridge University Press

4.3 Method

Like most variationist data, and natural language data in general, our dataset comes with traditional problems with regard to quantitative analysis. There are frequency fluctuations in terms of instances per informant and singletons; predictors and predictor levels are heterogeneously distributed and interrelated. All these problems have been previously described in the literature (for a compact overview see Tagliamonte & Baayen 2012:142–143). This reality sets boundaries for the quantitative methodology that can be applied in analysing the data. Although logistic mixed-effects modelling is frequently applied in variationist data analysis, we have to neglect this here as the assumptions for these modelling techniques are not met by our data (see also Tagliamonte & Baayen 2012:146, 161).

To overcome modelling restrictions posed by our data, we have chosen two non-parametric classification methods for our analyses: recursive partitioning tree models (Hothorn, Hornik & Zeileis 2006) and random forests (Breiman 2001, Strobl, Malley & Tutz 2009).

Recursive partitioning in a conditional inference framework (Hothorn et al. 2006) is, in simple terms, a recursive binary splitting of the data. The algorithm makes binary splits, deciding locally which variables best classify the data. When there are no significant factors left for splitting, the algorithm stops. The advantage of this method is that it enables us to identify interactions between variables and allows straightforward visualization to capture these interactions. Results are represented in a tree-like graph. We use the method specifically for these purposes: to visualize our data and to identify potential interactions between coded variables.

The primary method we apply is the random forests method (Breiman 2001) which complements the conditional inference trees. Random forests construct a large number of conditional inference trees and then select the variables that classify the data best. The method makes it possible to measure the relative importance of the variables included in the model. During the random permuting of the predictor variable, it measures the difference in prediction accuracy before and after permutation of the predictor, and thus measures the extent to which the model is weaker without the contribution of the predictor (Strobl et al. 2009). Random forests work well in situations with a relatively small number of observations and a large number of predictors (Tagliamonte & Baayen 2012), which is also obviously the case with the current dataset. Random forests have previously been successfully applied in linguistic studies (e.g. Tagliamonte & Baayen 2012, Baayen et al. 2013, Kyröläinen 2013).

All the computations were performed with program R (R Development CoreTeam 2013); we used the package *party* for both conditional inference tree (Hothorn et al. 2006) and random forests (Hothorn et al. 2006, Strobl et al. 2007, Strobl et al. 2008) analyses.

https://doi.org/10.1017/S0332586517000191 Published online by Cambridge University Press

For supplementary visualization purposes we use polygon gradiency maps, primarily to display the frequency data (Uiboaed 2016). On these maps, darker areas indicate higher frequencies and lighter areas lower ones. The frequency data were normalized on the basis of average file size in the corpus.

5. RESULTS

5.1 Lexical variation: **Tarvis** *or* **vaja**

In Estonian grammars *vaja* and *tarvis* (*olema*) have been always described jointly, as having the same functions and meanings. In this section, we show that these two modal adverbs have different origins and different distribution in the dialects which can be explained by local language contacts.

According to the Estonian etymological dictionary (Metsmägi, Sedrik & Soosaar 2012), *tarvis* is a Proto-Germanic loanword, also found in other Balto-Finnic languages: Finnish, Veps, Votic, Ingrian, Livonian and Karelian. *Vaja* is considered to be an old Finno-Ugric stem, also found in Livonian 'need', Votic 'need; missing', Finnish *vajaa* 'incomplete', Ingrian 'need, missing', Lude, Veps 'incomplete'. As a modal adverb, *vaja* is used in addition to Estonian, also in Votic (where the use of *tarvis* and *vaja* varies, according to the data of CED) and Livonian (Viitso 2014), i.e. southern Balto-Finnic languages.

According to the dictionary of contemporary Standard Estonian, *vaja* is approximately 3.5 times more frequent than *tarvis* (see Kaalep & Muischnek 2002). In contemporary spoken Estonian the same tendency is even more visible: *vaja* dominates clearly over *tarvis*, being used in almost 90% of the occurrences of *tarvis*/*vaja* constructions (Lindström, Uiboaed, Vihman 2014, Lindström & Vihman 2017). There are no dialectal data where the distribution of *tarvis* and *vaja* are comprehensively described, but the Estonian dialect dictionary VMS (Haak et al. 1989) contains information as to sub-dialects where these occurrences have been attested. Figures 2 and 3 were retrieved from the VMS web version (http://www.eki.ee/dict/vms/) and illustrate the distribution of *tarvis* and *vaja* in sub-dialects.

The maps in Figures 2 and 3 mark the locations (sub-dialects) where the corresponding data were collected, but the graphs do not convey frequency information; thus this kind of atlas data do not show whether the adverb occurred at that location once or a thousand times. Comparison of these maps does not reveal any crucial difference between the distribution of *tarvis* and *vaja* across Estonian dialects; rather, both adverbs seem to occur quite homogeneously all over the region.

With corpus data, in addition to location information, we are also able to retrieve the frequencies of the observed forms. Table 1 and Figure 4 demonstrate the frequency differences between dialects.

https://doi.org/10.1017/S0332586517000191 Published online by Cambridge University Press

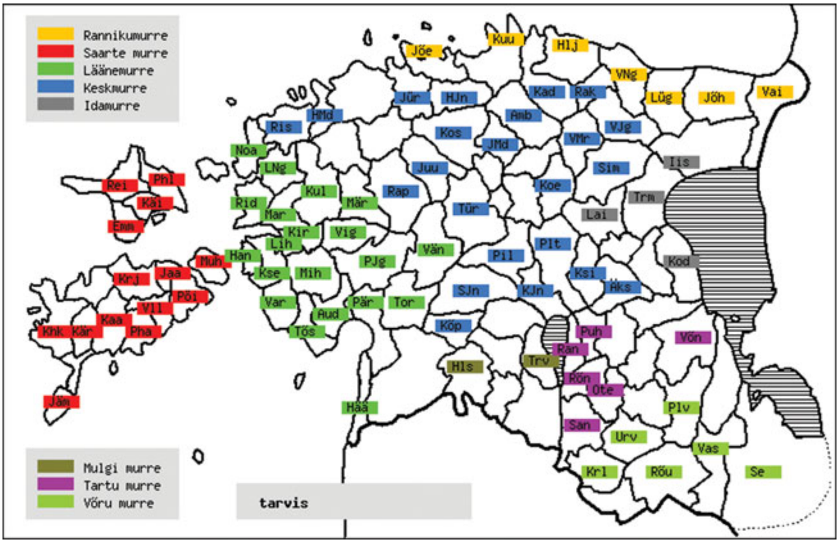

Figure 2. (Colour online) The spread of *tarvis* in Estonian dialects and sub-dialects (different colours mark different dialects; abreviations label sub-dialects).

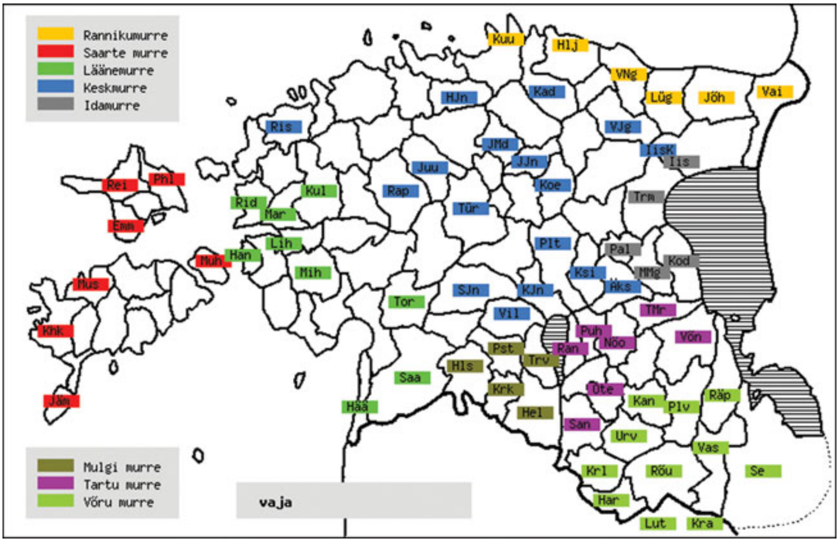

Figure 3. (Colour online) The spread of *vaja* in Estonian dialects and sub-dialects (different colours mark different dialects; abreviations label sub-dialects).

Figure 4 visualizes the proportions of *tarvis* constructions out of all constructions. The proportions were calculated on the basis of the frequencies of ‘need’-constructions, i.e. how many of these constructions are formed with *vaja* or *tarvis* respectively. Darker areas indicate higher proportions of *tarvis*,

https://doi.org/10.1017/S0332586517000191 Published online by Cambridge University Press

Dialect	*tarvis*	*vaja*	Total
Coastal	16	5	21
Eastern	9	33	42
Insular	50	1	51
Mid	37	7	44
Mulgi	3	16	19
Northeastern	16	7	23
Seto	0	38	38
Tartu	5	20	25
Western	68	11	79
Võru	0	13	13
Total	204	151	355

Table 1. Frequencies of *tarvis* and *vaja*.

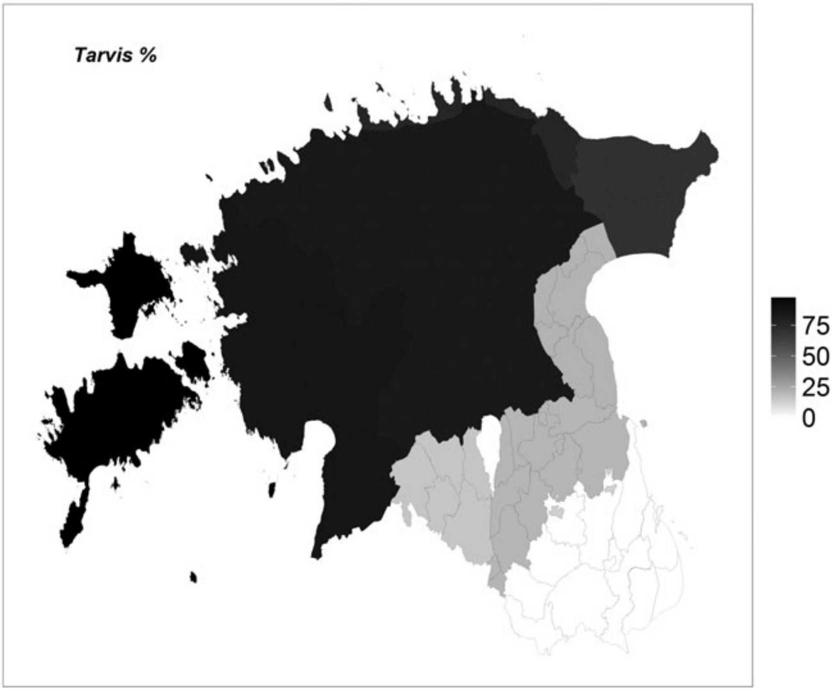

Figure 4. The proportions of *tarvis* and *vaja* out of all instances observed. Darker areas indicate a higher proportion of *tarvis*, lighter ones a preference for *vaja*.

lighter areas vice versa: lower proportions of *tarvis* and thus higher proportions of *vaja*.

According to Figure 4, the use of *tarvis* decreases and that of *vaja* increases from north-west to south-east. South-Estonian dialects mainly (or only, as in Võru

https://doi.org/10.1017/S0332586517000191 Published online by Cambridge University Press

and Seto) use *vaja*, which is also dominant in the Eastern dialect, traditionally classified as belonging to the North-Estonian dialect group. The preference for *vaja* in the southern and eastern dialects may, on the one hand, be explained by old lexical developments in Southern Balto-Finnic languages, especially in South Estonian and Livonian; on the other hand, its use has probably been influenced by Latvian, which is a close neighbor of the South-Estonian dialects and uses the same stem as a modal verb (*vajadzēt* 'need') in similar functions. Interestingly, the Latvian verb *vajadzēt* has been found to be a Finnic loanword which was subsequently verbalized (Seržant & Bjarnadóttir 2014); thus the influence seems to be mutual: Latvian first borrowed the stem from Finnic (either Estonian or Livonian) and verbalized it, while the later Latvian *vajadzēt* has reinforced the preference for *vaja* especially in the South Estonian dialects.

In western dialects (Western, Insular, also Mid, appearing as a more homogenous group in the current analyses) the use of *tarvis* dominates over *vaja*, being almost the only option in the Insular dialect. Coastal and North-Eastern dialects, on the one hand, show a preference for *tarvis*; on the other hand, the frequencies indicate that the use of the adverb inclines more to variation.

Merely from observing frequencies, it is evident that lexical item (*tarvis*/*vaja*) of the construction varies across dialects and different dialects tend to prefer either one of them. Nevertheless, in most dialects both modal adverbs can occur in the construction. In order to explore how other predictors potentially contribute to the choice between modal adverb and how important these predictors are when it comes to the choosing one of the modal adverbs we applied two methods: recursive partitioning tree and random forests analysis. The first method is better for classifying sub-groups of the predictors whereas the second, random forests, measures the importance of individual variables. Every sub-analysis first presents the results of the recursive partitioning tree followed by the results of random forests analysis.

We conducted two different analyses on slightly different datasets: with the individual variable excluded and included. The individual variable is nested with the dialect variable. Our binary dependent variable in the first model is the lexical item (adverb *tarvis* or *vaja*); the purpose is to explore which predictors best explain the use of *tarvis* or *vaja*. The predictors in the model are dialect, main verb, complement, experiencer case-marking, polarity and tense. Figures 5 and 6 show the recursive partitioning tree results for lexical variation.[2] The data are split by the predictors which most strongly influence the choice between two lexical items. The first split is the most important one, and the next splits are based on the significance of other predictors. Divisions are made until there are no further significant sub-classifications. As the raw data already suggested, the most powerful explanatory predictor in both analyses is the dialect. Both trees place the Coastal, Insular, Mid, Northeastern and Western dialects in one group (clearly the *tarvis* group), and the Eastern, Mulgi, Seto, Tartu and Võru dialects in another (the *vaja* group).

https://doi.org/10.1017/S0332586517000191 Published online by Cambridge University Press

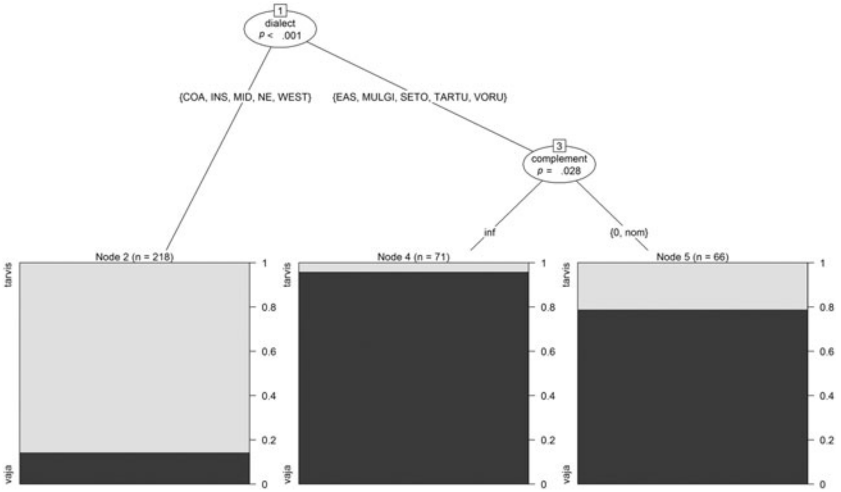

Figure 5. Recursive partitioning tree (lexical variation between *tarvis* and *vaja*) without the individual variable.

The results are very similar; we are clearly not primarily dealing with individual variation.

Within the *vaja* group (containing Eastern, Mulgi, Seto, Tartu and Võru dialects), however, a separate split is made on the complement (in both analyses); in other words, the type of complement predicts the choice between *tarvis* and *vaja* in this dialect group. If a nominal or zero complement is expressed, *tarvis* is more likely to be used, while *vaja* is used more often in infinitival (modal) constructions. Thus Southern and Eastern dialects, where *vaja* dominates, use it more often as a modal predicate, illustrated in (8).

(8) perenainõ nakka-ss et üttel' et kar'ussõ-lõ vaja süvvä
farmwife start-3SG that say.PST.3SG that shepherd-ALL need eat.INF
anda (Võru)
give.INF
'the farmwife said that (somebody) has to give the herder something to eat'

The second, *tarvis* group (Coastal, Insular, Mid, Northeastern and Western dialect) also exhibits large inter-individual variation.

We applied random forests analyses on the same datasets. The left-hand panel in Figure 7 shows the variable importance plot when the individual was excluded from the analysis (C = 0.95),[3] the right-hand panel when it was included (C = 0.98). The figure illustrates the important variables in both analyses: variables on the right side of the light gray vertical line are important in discriminating between the two responses. Both analyses reveal that dialect is definitely the most important

https://doi.org/10.1017/S0332586517000191 Published online by Cambridge University Press

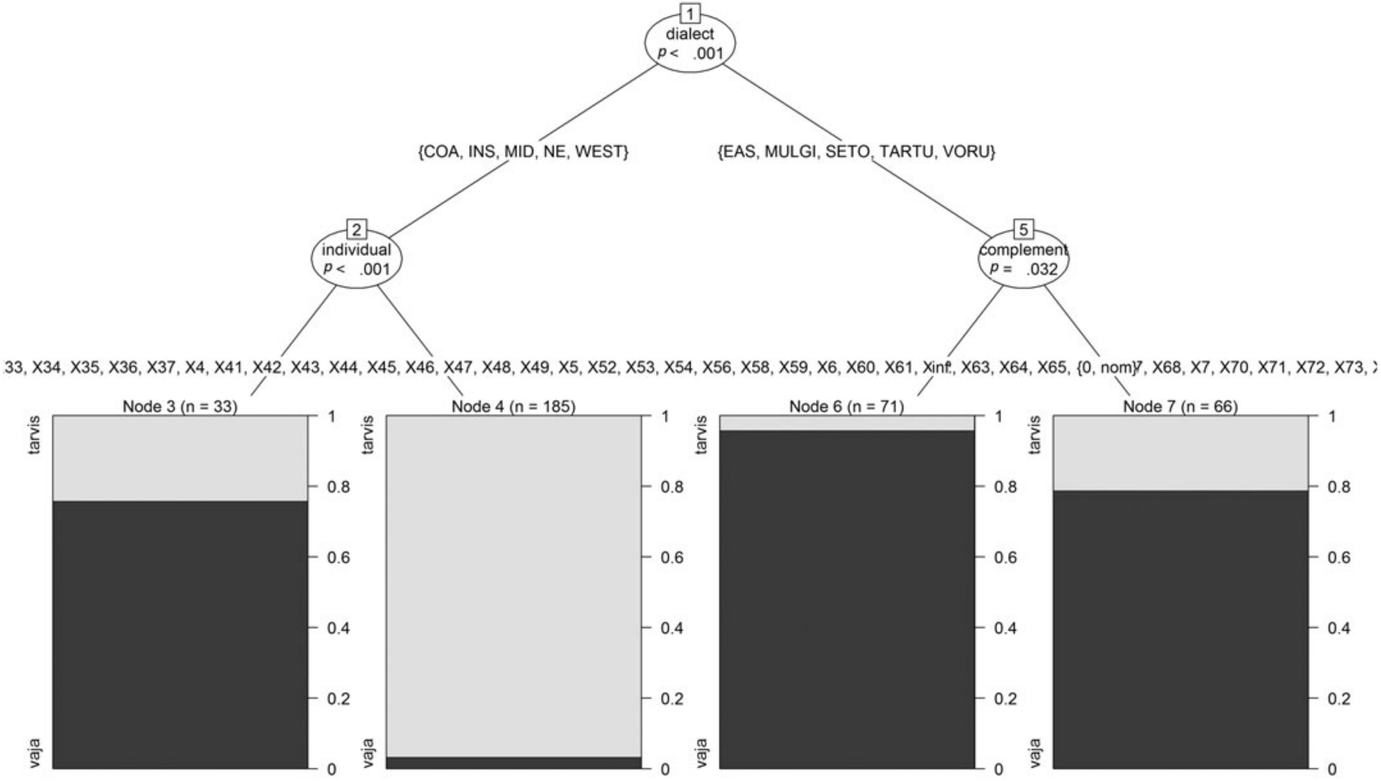

Figure 6. Recursive partitioning tree (lexical variation between *tarvis* and *vaja*) with individual variable.

https://doi.org/10.1017/S0332586517000191 Published online by Cambridge University Press

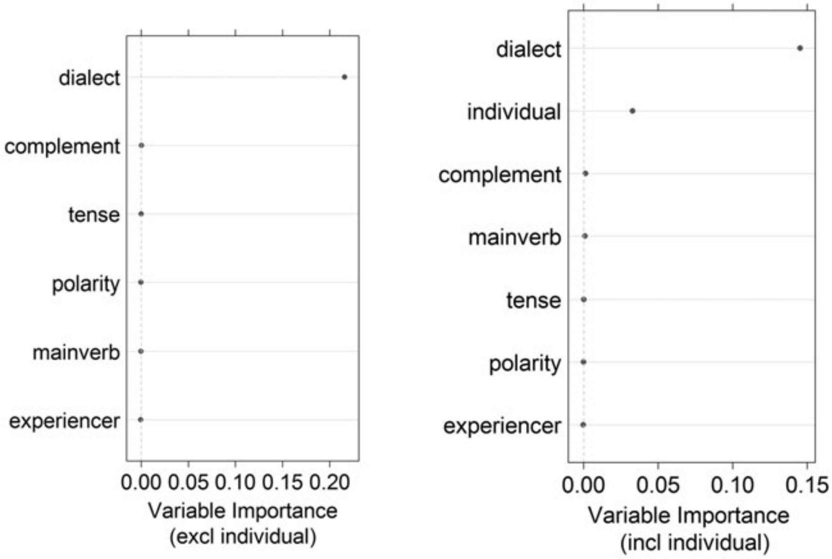

Figure 7. Variable importance (lexical variation between *tarvis* and *vaja*).

variable in the choice between *tarvis* or *vaja*. Surprisingly, in the second analyses the individual is a far less important variable than dialect. This gives a strong indication as to the influence of geography in deciding between *tarvis* or *vaja*.

Thus the distribution of *tarvis* and *vaja* is clearly determined by dialect; of the other factors, only complement type has some importance in the recursive partitioning tree models. The South-Estonian dialects (Mulgi, Tartu, Võru, Seto) and the Eastern dialect prefer *vaja*, while other North-Estonian dialects more often use *tarvis*.

At this point we can draw one methodological inference, based on the frequency data. Although our corpus sample is relatively small, differences between the dialects are clearly distinguishable and illustrate the features that are more (or less) typical of particular dialects. This is more fine-grained information than that provided by simple dictionary data with distributional statements. For practical reasons, we aggregated over relatively large areas (dialects). Given sufficient sampling data from smaller areas (sub-dialects) to yield frequency-based conclusions, we would very likely see a map with smooth transitions from South-East to North-West.

5.2 Variation in copula use

Another highly variable part of the construction across dialects is the copula verb, which is typically *olema* 'be', as in (3) (repeated below), but motion verbs used similarly to the copula *tulema* 'come', as in (9), and *minema* 'go', as in (10), can also appear in this position. Frequencies of copulas and copula ellipsis in different dialects can be observed in Table 2.

https://doi.org/10.1017/S0332586517000191 Published online by Cambridge University Press

Dialect	Verb ellipsis	*minema* ‘go’	*olema* ‘be’	*tulema* ‘come’	Total
Eastern	18	0	23	1	43
Mid	9	3	32	0	44
Northeastern	3	1	18	1	23
Western	11	4	61	3	79
Mulgi	2	0	11	0	19
Coastal	4	0	17	0	21
Insular	2	4	45	0	51
Seto	27	0	11	0	38
Tartu	8	0	17	0	25
Võru	3	0	10	0	13
Total	87	12	252	5	355

Table 2. Frequencies of copulas and copula ellipsis.

(3) kui su-lle ädäste tarviss omm siss mine ja
if you-ALL so_much need be.3SG then go.IMP and
võtta tu raha (Tartu)
take.IMP this money
‘if you need (it) so much, go and take this money’

(9) ja siss tul’l-i nagu rohkem raha tarviss (Western)
and then come-PST.3SG like more money.PRT need
‘and then we needed more money’

(10) uks on käemabaik sealt lähe-b ühtebuhku
door be.3SG place.to.go there go-3SG often
tarvis käia (Mid)
need go.INF
‘a door is a place to go in and out, where you need to do that all the time’

Uses of *tulema* ‘come’ and *minema* ‘go’ are infrequent in these constructions and tend to occur mainly with the nominal complement. (Example (10) is exceptional in this respect, as it takes the infinitival complement.) The verb *tulema* occurred in our data only five times and *minema* twelve times; thus we cannot say anything conclusive about the geographical distribution of these verbs, but we can report that none of them occurred in South Estonian dialects.

A more interesting picture is revealed when we look at the ellipsis – *0* – of the copula in (11).

(11) ja maa mõtle-n miss ma tii-n tämä-ga, *0* vaja kedägi
and I think-1SG what I do-1SG(s)he-COM need something.PRT
panna ta-lle sinna kot’ti (Eastern)
put.INF (s)he-ALL there bag.ILL
‘and I think about what I shall do with him, (I) need to put something in his bag’

https://doi.org/10.1017/S0332586517000191 Published online by Cambridge University Press

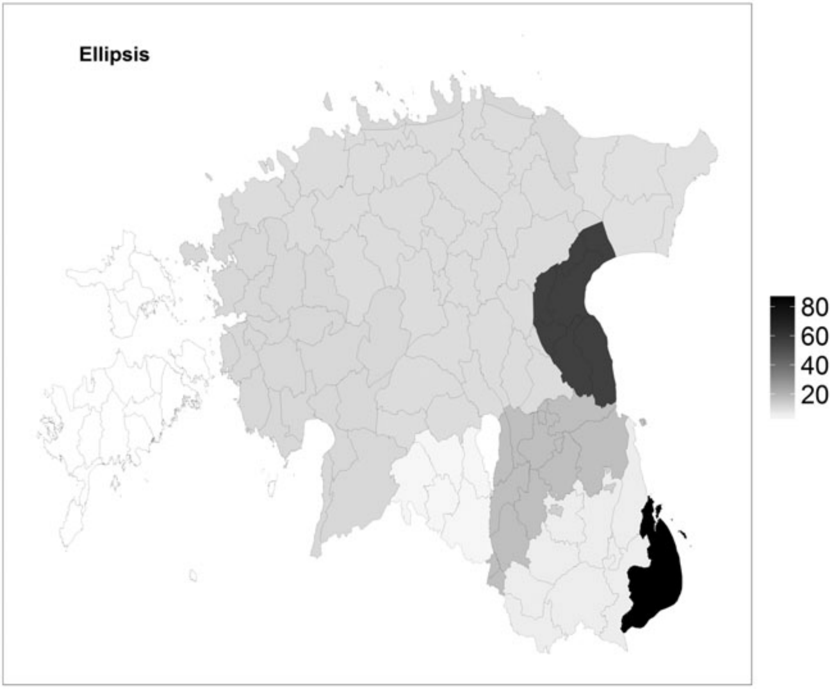

Figure 8. Frequency distribution of ellipsis of the copula.

Ellipsis of the copula verb is characteristic of Russian, a contact language of Estonian. Hansen (2014:104) states that in Russian, the copula is widely omitted in modal constructions with modal adverbs. Kehayov (2009:129) has studied many of the contexts where the ellipsis of the copula occurs in Estonian, and found that one such context consists of clauses with *tarvis/vaja*. He accounts for the ellipsis in terms of Russian influence. Kehayov & Torn-Leesik (2009) also claim that ellipsis of the copula verb with the modal predicates *on vaja/tarvis* is very common in most of the eastern and southern Balto-Finnic languages (Veps, Karelian, Votic, Livonian), which can again be explained by Russian contacts.

Figure 8 shows the gradience map of verb ellipsis in *tarvis/vaja* constructions. The frequencies are normalized on the basis of the mean file size in the corpus.

It may be noted that ellipsis of the verb occurs more frequently in the Eastern and Seto dialects, being clearly the most usual pattern in Seto. Thus it can be confirmed that ellipsis is especially common in those dialects that have had long direct contact with Russian. The Seto dialect is spoken on the border between Estonia and Russia, while the Eastern dialect has had long-lasting contact with Russian Old Believers,

https://doi.org/10.1017/S0332586517000191 Published online by Cambridge University Press

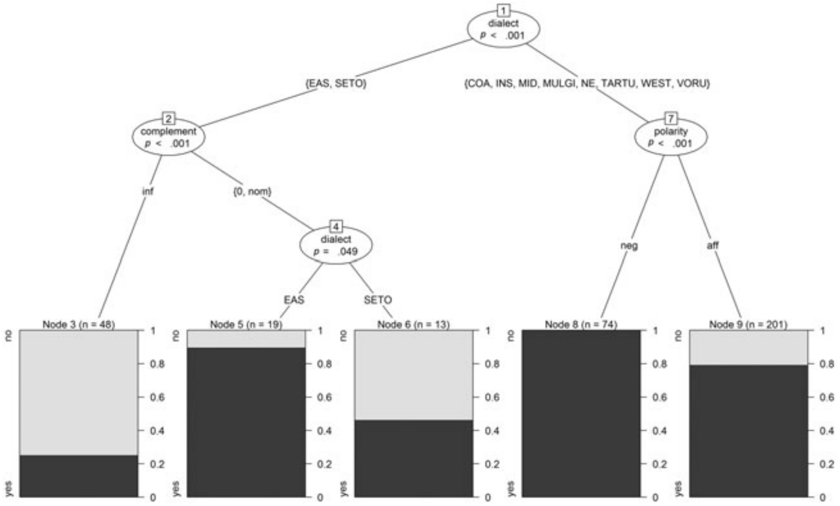

Figure 9. Recursive partitioning tree (copula ellipsis) without individual variable.

who have been living on the shore of Lake Peipus since the seventeenth century (Čekmonas 2001). In western dialects, by contrast, ellipsis occurs rarely; for example, in the Insular dialect only about 4% of clauses with *tarvis*/*vaja* occur without the copula verb. Our data thus support Kehayov's (2009:139) findings: that ellipsis in *tarvis*/*vaja* constructions is influenced by the analogous phenomenon in Russian, and that it occurs more often in areas of close contact with Russian.

However, the copula is needed for expressing certain grammatical categories; thus there are other factors, such as polarity and tense, which may also affect the occurrence of the copula (Kehayov 2009:143). In order to investigate which predictors influence the presence or absence of the copula (binary dependent variable in the following two models), we included the following predictors in our tree model: dialect, complement, polarity, tense, case-marking of experiencer argument, case-marking of nominal complement (if present). Figure 9 shows the results of the copula ellipsis analysis without the *individual* variable.[4] Analyses with the *individual* variable included produced the same classification results; thus we do not present the tree here.

Both analyses provided very similar results. Predictor *dialect* was the most important predictor for presence vs. absence of a copula verb. In the group consisting of the Coastal, Insular, Mid, Mulgi, Northeastern, Tartu, Western and Võru dialects, the split is made by the predictor *polarity*, while *tense* did not emerge as a statistically significant predictor. Thus ellipsis (*0*) may also occur in referring to past time, as in (12), where the clause containing the necessive construction occurs in a past context.

https://doi.org/10.1017/S0332586517000191 Published online by Cambridge University Press

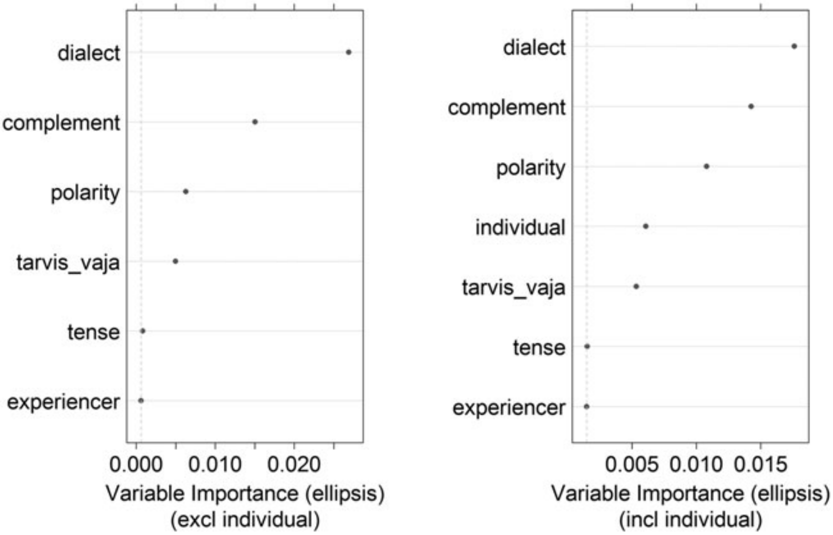

Figure 10. Variable importance (copula ellipsis).

(12) jahh sa-i (.) sa-i sie peso (.) ülesse pan-dud kõik
yeah get-PST.3SG get-PST.3SG this laundry up put-PPP all
juo ja si- siis tul-i-n maha ja (.)
already and the- then come-PST-1SG down and
noh siis viel õl-i tüö-d (.)
PART then more be-PST.3SG work-PRT
siis tarvis luoma-d tasu-da (.)
then need animal-PL tend-INF
siis tasu-si-n luoma-d ärä kõik ja (North-Eastern)
then tend-PST-1SG animal-PL away all and
'yeah, the laundry was (temporarily) closed then I came down and there was some work, then I tended the animals and'

In the second dialect group, consisting of the Seto and Eastern dialects, the occurrence of the infinitival complement is a statistically important predictor with regard to verb ellipsis. Interestingly, if the infinitive occurs as a complement, the copula tends to be omitted (as it was also in example (12) from North-Eastern). The copula is more likely to be expressed if the complement is nominal or is omitted.

To test the importance of various factors in determining verb ellipsis, we conducted random forests analyses on these two datasets. The results are displayed in Figure 10. Again the first model (left-hand panel) excludes the individual variable (C = 0.90) while the second (right-hand panel) includes (C = 0.95) it. In both models, the most important variable is the dialect, as was also clear from the tree model. But complement type, polarity, and lexical item (*tarvis*/*vaja*) also gain importance in both models. The second model (right-hand panel) also reveals some individual variation, but it is the far less important variable, and it is clear that in the constructions observed, copula ellipsis is more strongly driven language-internally.

To conclude this section, ellipsis of the copula verb shows clear dialectal divergence: ellipsis is most common in the Seto and Eastern dialects, which are direct contact areas of Russian. It occurs most rarely in the westernmost part of the region, in the Insular dialect. Within dialect groups, ellipsis is dependent on language-internal factors, such as the occurrence of the infinitival complement (with more frequent copula ellipsis) and polarity (with more frequent use of the copula).

5.3 Expressing the experiencer

The experiencer argument in Estonian 'need'-constructions is case-marked with either the adessive (13) or the allative (14).

(13)	et	nooq	minu-l	olõ-ss	nüüd	kohta	vaja	(Tartu)
	that	PART	*I*-ADE	*be*-CND.3SG	*now*	*place*.PRT	*need*	

'I need a position now'

(14)	edespädi	tä-lle	tarviss	eij	ole	ollu	(Tartu)
	later	*(s)he*-ALL	*need*	*not*	*be*.CNG	*be*.APP	

'later (s)he has not needed (it)'

The omission of the experiencer (zero marking) is possible as well. Zero marking includes on the one hand an indefinite 'zero-person' (general or indefinite uses): in (15) below, the preceding and following clause contain an impersonal verb form, typically used for indefinite reference (see Torn-Leesik & Vihman 2010); on the other hand, zero anaphora, in which reference is implied by the context in (16). These two zero types, however, are not always clearly distinguishable, and reference resolution emerges from the context. (See also Lindström & Vihman 2017.)

(15)	need	vii-di	koa	ves'ki-lle	kudas	kudas	tarvis	ol-i
	these.PL	*bring*-IPS.PST	*too*	*mill*-ALL	*how*	*how*	*need*	*be*-PST.3SG
	võe-tti	sialt	võe-tti	sieme	(Mid)			
	take-IPS.PST	*there*	*take*-IPS.PST	*seed*				

'these were brought to the mill when needed, (they) took seeds from there'

(16)	kolm	aasta-d	maa	käi-si-n	ja	rohkem	pöl-d
	three	*year*-PRT	*I*	*go*-PST-*1*SG	*and*	*more*	NEG.*be*-PST.PTCL
	tarvis					(Mid)	
	need						

'I went (there) for three years and (I) didn't need any more'

Experiencers marked with the adessive or the allative are common in Estonian in various constructions, such as modal constructions with the verb *tulema* 'come; have to' (Penjam 2006) and *tarvitsema* 'need' (Penjam 2011), the so-called possessive perfect construction (see Lindström & Tragel 2010), in some experiential constructions, and many others. In these constructions, they share certain subject properties: for instance, they are used clause-initially, they refer to human referents, they are typically pronominal, and they share certain syntactic properties characteristic of prototypical subjects (Keenan 1976; for Estonian, see Metslang 2013). Thus the experiencer of *tarvis/vaja* constructions is a subject-like argument, although case-marked differently from the prototypical (nominative) subject. The abundant use of non-nominative, dative-like experiencers in experiential constructions is considered to be a common feature of the Eastern Circum-Baltic languages, including Estonian, Russian and Latvian (Seržant 2015b). Hansen (2014) has drawn similar conclusions concerning modal constructions: a non-canonical marking of subject(-like) arguments in modal constructions is characteristic of the East Slavonic, Baltic and Balto-Finnic languages (see also Kehayov & Torn-Leesik 2009), and may thus be an areal feature. He also mentions that non-nominative modal subjects may often be omitted, but that the conditions of the omission in these languages are unclear (Hansen 2014). Non-nominative marking of modal subjects is found also in Finnish (the experiencer is marked mainly by genitive, but in colloquial speech and dialects also in adessive), and in Finnish, too, they are often omitted (Laitinen 1992, ISK:1291).

In this section, we examine the constraints on the choice between zero and overt marking of the experiencer. While the question of choice between adessive and allative marking of the experiencer is of interest, we do not analyse it here because the data are insufficient for quantitative analysis. We can only conclude that use of the allative was limited: it occurred 15 times in our data, and mostly in the eastern part of Estonia (see Table 3). Moreover, the use of allative seems to be syntactically restricted: it occurs only with a nominal complement, there was only one exception with an infinitival complement. The use of allative experiencers in eastern dialects can be explained by the influence of Russian: the semantics of dative case in Russian is closer to allative than to adessive as it has some directionality in its meaning, and it is used also for marking the recipient (Janda 2008), similarly to Estonian allative case. On the other hand, in western dialects the adessive and allative are often collapsed: allative ending –(*l*)*le* has been shortened to -(*l*)*l*, which is used also for adessive; eastern dialects, on the contrary, have preserved the distinction.

Table 3 shows that the dominating pattern is to omit the experiencer: 268 instances (75%) out of the 355 occur without an explicit experiencer argument. Figure 11 shows this in a visual form: darker areas indicate higher proportions of areas where the experiencer is not expressed.

	Occurrence of the experiencer			
		Yes		
Dialect	No	Adessive	All	Total
Coastal	16 (76%)	5	0	21
Eastern	29 (69%)	11	2	43
Insular	38 (75%)	13	0	51
Mid	37 (84%)	7	0	44
Mulgi	12 (63%)	7	0	19
Northeastern	15 (65%)	7	1	23
Seto	32 (84%)	3	3	38
Tartu	15 (60%)	3	7	25
Western	63 (80%)	15	1	79
Võru	11 (85%)	1	1	13
Total	268 (75%)	72	15	355

Table 3. Frequencies of experiencer-marking.

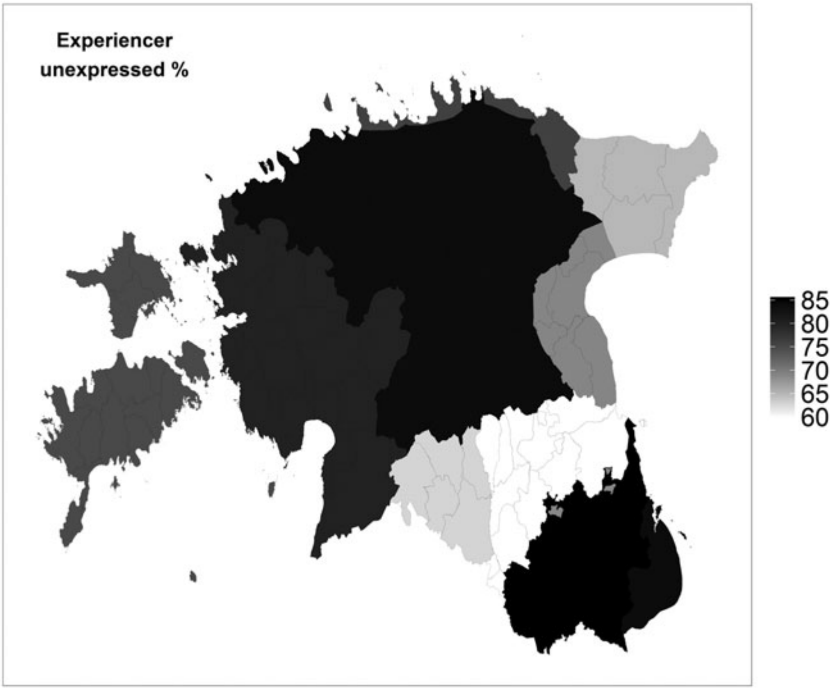

Figure 11. Proportions of experiencer omission.

https://doi.org/10.1017/S0332586517000191 Published online by Cambridge University Press

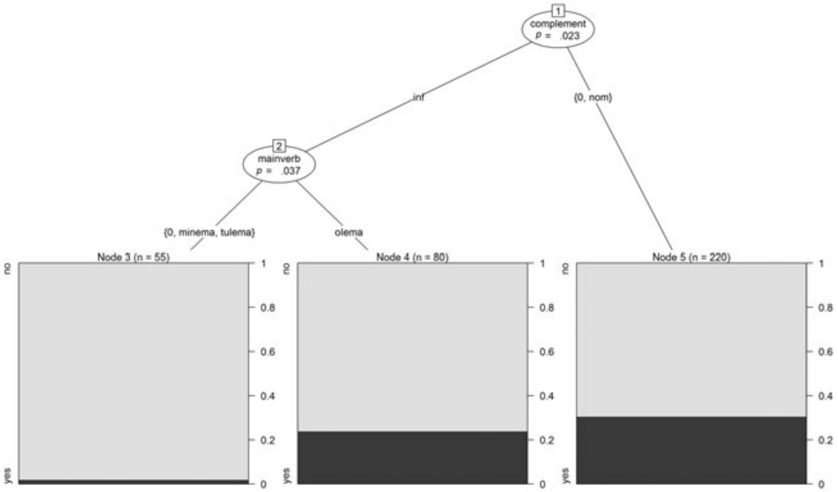

Figure 12. Recursive partitioning tree (experiencer marking) without individual variable.

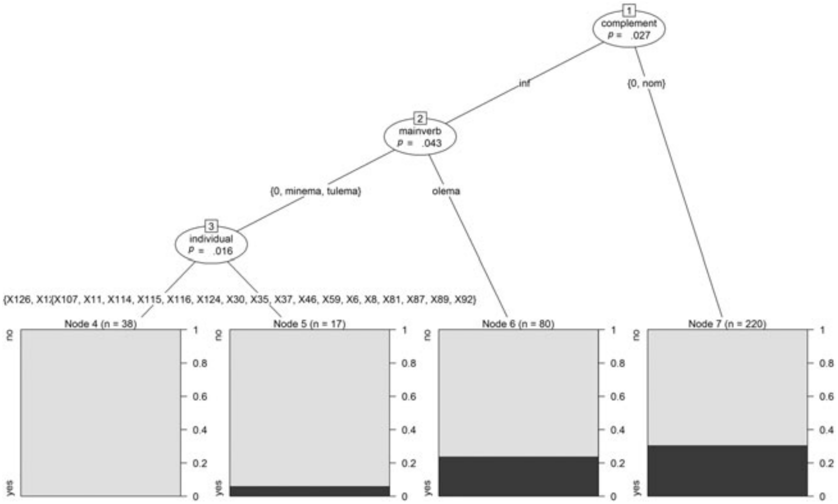

Figure 13. Recursive partitioning tree (experiencer marking) with individual variable.

Again we ran two different analyses where our binary dependent variable was the occurrence of the experiencer argument. Figures 12 and 13 show the results of the recursive partitioning tree.[5]

Both analyses reveal that dialect is not an important predictor when it comes to experiencer-marking. Most significant are language-internal factors, above all the complement type. When the infinitival complement occurs, the copula (predictor *mainverb*) emerges as a significant factor; when the copula *olema* ‘to be’ is

https://doi.org/10.1017/S0332586517000191 Published online by Cambridge University Press

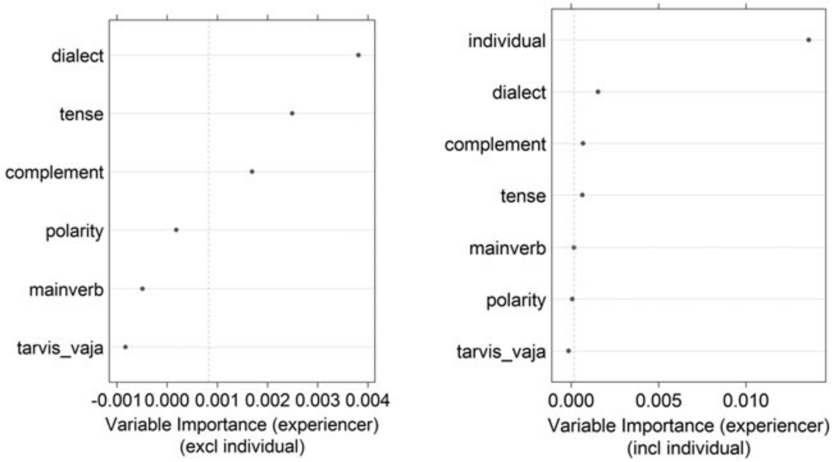

Figure 14. Variable importance (marking of the experiencer).

present, the experiencer argument is more likely to occur. Thus the presence of the infinitival complement and absence of the copula reduces the probability of overt expression of the experiencer. The importance of the predictor *complement* indicates that there is a difference between the two main ‘need’-constructions: with the infinitival complement (i.e. in the modal ‘need’-construction, example (17) below, the experiencer is less often overtly expressed than with nominal, seen in (18), and zero complement. The presence or absence of the experiencer is thus syntactically determined.

(17)	tütre-d	ol-i-d	kodu,	noo	aga	kodu	ol-i	tar’vis	kaa
	daughter-PL	*be-PST-3PL*	*home*	*but*	*but*	*home*	*be-PST.3SG*	*need*	*also*
	olla	(Mid)							
	be.INF								

‘the daughters were at home, but that’s also where they had to be’

(18)	minu-l	eij	õl-d	se-da	tarvis	(North-Eastern)
	I-ADE	*not*	*be-PST.PTCL*	*it-PRT*	*need*	

‘I didn’t need it’

The results of the random forests analyses are shown in Figure 14. The left-hand panel shows the importance of the variable for the analyses when the individual was excluded (C = 0.83), the right-hand panel shows the results with the individual included (C = 0.96).

The index of the concordance shows that the second model better explains our data, so we see substantial individual variation. Comparison of the two panels indicates that when the predictor *individual* is taken into account, all other variables

https://doi.org/10.1017/S0332586517000191 Published online by Cambridge University Press

are clearly less important in the model; thus, while complement, dialect and tense play some role in the omission of the experiencer, individual differences determine the expression of the experiencer argument, above all.

In experiencer-marking, the two methods used (recursive partitioning tree and random forest) lead to different results, which at first glance are hard to explain. However, the recursive partitioning tree better reveals the distribution and interactions within the data, while random forests measure the role of each predictor separately; thus we have to take both analyses into account. To sum up: on the one hand syntactic factors (complement type and copula) affect the use of the experiencer, on the other hand there are considerable differences among individuals. Dialect differences seem to be less important than individual ones, as shown in Figure 13 above. Hence, there is no reason to assume that the expression or omission of the experiencer is geographically based, and neither can it be explained by local language contacts. All the dialects show high rates of experiencer omission (60–85%), without clear regional differences. Therefore, we can assume that the omission of the experiencer is an old feature of Estonian and not a late contact-induced phenomenon, although the same is found in Slavic and Baltic languages as well. The rate of omission seems to be related to syntactic factors, but also varies between the individuals.

However, the differences between the results of the two methods may also be a sign that there is something else going on, not taken into account so far. Another study (Lindström & Vihman 2017) reveals that in the same constructions in standard Estonian, participant-internal or -external modality has a crucial role, with participant-internal modality clearly increasing the probability of overt expression of the experiencer. In the present study modality is not analysed in detail, so we cannot further elaborate upon the idea of the effect of modality type on experiencer ellipsis in dialects.

6. DISCUSSION AND CONCLUSION

In this paper, we examined two ‘need’-constructions, consisting of the adverb (*tarvis*/*vaja*) + nominal/infinitival complement. Although it might seem that we are dealing with relatively idiosyncratic phenomena, these constructions enable us to illuminate the complexity of areal syntactic variation from a number of perspectives. These constructions showed lexical variability (modal adverb *tarvis*/*vaja*), construction-internal variability (nominal vs. infinitival complement, copula verb and ellipsis variation, experiencer-marking), and both geographical and individual variation.

Our first analyses (Section 5.1) revealed that the lexical component of the construction has a clear dialectal distribution: *tarvis* appears more often in the western part of the country, while *vaja* is clearly characteristic of southern and eastern dialects. Omission of the copula verb is more characteristic of dialects that have had long-lasting contact with Russian. The results of the quantitative analyses indicate

https://doi.org/10.1017/S0332586517000191 Published online by Cambridge University Press

that in addition to dialectal differences, language-internal factors play a role as well; in particular complement type and polarity have a considerable impact on copula omission. Interestingly, individual differences are much less important here than in the other analyses, suggesting that omission of the copula is strongly driven by dialectal and language-internal factors.

In marking the experiencer argument, in contrast, the variation seems to be highly individual; individual differences are more important than dialectal ones. Language-internal factors, however, also play an important role, especially complement type: constructions with a nominal complement tend to express the experiencer overtly more often than modal (infinitival) constructions. The analyses also indicated that including the individual variable definitely results in better statistical models and data variability is better explained. Interestingly, the variable *individual* clearly was not the most important predictor in all our analyses. This confirms that in 'need'-constructions, individual variation is secondary and the constructions exhibit very strong dialectal variation. Moreover, different components of the constructions are very strongly connected to language-internal factors additionally to dialectal ones.

In this study, we analysed variation in complex constructions investigating each item in the constructions separately. We excluded from the analysis that part of the variation that took place at a more local level (e.g. case-marking of the nominal complement, variation between adessive and allative marking of the experiencer); the scope of the research was limited by the lack of data for quantitative methods due to the small size of the corpus.

The main purpose of the study was to ascertain what part of the variation in these constructions is geographically driven and what is determined by language-internal factors. In 'need'-constructions, the role of the contact was the most evident in ellipsis of the verb, as the ellipsis occurred clearly more often in Eastern and Seto dialects that have had long and direct contact with Russian. In the choice of modal adverb, the inter-dialectal differences were obvious, but it is not clear whether the preference for *vaja* is related to specific lexical developments in Southern Balto-Finnic languages in general, or rather a result of mutual influences between South Estonian and Latvian. The role of language contact was the least evident in omission of the experiencer argument: there were no clear dialectal preferences, the omission was usual in all dialects. The experiencer can be omitted also in Finnish, which has had less recent contact with Russian and Latvian, indicating that omission is usual in Balto-Finnic languages, and it is not a contact-induced phenomenon but rather characteristic to all Balto-Finnic (or even Uralic) languages.

The study has provided new insights into areal syntactic variation and has highlighted the complexity of this kind of variation. Not merely the geographical spread of a particular linguistic item or form, but also the individual parts of a complex construction should be kept in mind; their spread and interaction vary both geographically and language-internally.

ABBREVIATIONS

1, 2, 3	first, second, third person
ADE	adessive
ALL	allative
APP	active past participle
CED	Corpus of Estonian Dialects
CND	conditional
CNG	connegative
DAT	dative
F	feminine
ILL	illative
IMP	imperative
INF	infinitive
IPS	impersonal voice
M	masculine
NEG	negation
PART	particle
PL	plural
PPP	past passive participle
PRS	present tense
PRT	partitive
PST	past tense
PTCL	participle
SG	singular

ACKNOWLEDGEMENTS

This study was supported by the Estonian Research Council (project PUT90 'Estonian Dialect syntax') and by the (European Union) European Regional Development Fund (Centre of Excellence in Estonian Studies). We are grateful to three anonymous reviewers and editors of *NJL* for useful comments.

NOTES

1. Examples come from the Corpus of Estonian Dialects and follow the transcription of the corpus. As the corpus is a spoken corpus, it contains some special conventions for features of spoken language, e.g. (.) stands for a short pause, dash (e.g. *si*-) for unfinished words, and ' for palatilization.
2. First analysis is calculated with the formula: tarvis_vaja ~ dialect + mainverb + experiencer + complement + polarity + tense, the second with the formula tarvis_vaja ~ dialect + mainverb + experiencer + complement + polarity + tense + individual. The dependent

https://doi.org/10.1017/S0332586517000191 Published online by Cambridge University Press

variable is anteceds ~ and is followed by independent variables, each of these is separated with +.

3. We use index of concordance (C) (Harrell 2001:247) to estimate the goodness of the model: in the present case the value is the indication how well the model discrimantes between *tarvis* and *vaja* constructions. The value of C ranges between one and zero and C ≥ 0.8 is considerede to be a good performance (Tagliamonte & Baayen 2012: 156).
4. The first analysis: verb_is ~ dialect + tarvis_vaja + polarity + tense + experiencer + complement. The second analysis: verb_is ~ dialect + tarvis_vaja + polarity + tense + experiencer + complement + individual
5. The first analysis: exp_is ~ dialect + tarvis_vaja + mainverb + complement + polarity + tense. The second analysis: exp_is ~ dialect + tarvis_vaja + mainverb + complement + polarity + tense + individual

REFERENCES

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(3), 253–291.

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

Bucheli Berger, Claudia, Elvira Glaser & Guido Seiler. 2012. Is a syntactic dialectology possible? Contributions from Swiss German. In Andrea Ender, Adrian Leemann & Bernhard Wälchli (eds.), *Methods in Contemporary Linguistics* (Trends in Linguistics: Studies and Monographs 247), 93–120. Berlin & Boston, MA: De Gruyter Mouton.

Čekmonas, Valeriy. 2001. Russian varieties in the southeastern Baltic area: Rural dialects. In Dahl & Koptjevskaja-Tamm (eds.), vol. 1, 101–136.

Croft, William. 2001. *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford: Oxford University Press.

Dahl, Östen & Maria, Koptjevskaja-Tamm. (eds.). 2001. *Circum-Baltic Languages: Typology and Contact,* vol. 1: *Past and Present &* vol. 2: *Grammar and Typology* (Studies in Language Companion Series 54–55). Amsterdam & Philadelphia, PA: John Benjamins.

Erelt, Mati, Tiiu Erelt & Kristiina Ross. 2000. *Eesti keele käsiraamat* [Handbook of the Estonian language]. Tallinn: Eesti Keele Sihtasutus.

Erelt, Mati, Reet Kasik, Helle Metslang, Henno Rajandi, Kristiina Ross, Henn Saari, Kaja Tael & Silvi Vare. 1993. *Eesti keele grammatika II. Süntaks. Lisa: kiri* [Estonian reference grammar II: Syntax]. Tallinn: Eesti Teaduste Akadeemia Keele ja Kirjanduse Instituut.

Grieve, Jack. 2009. *A Corpus-based Regional Dialect Survey of Grammatical Variation in Written Standard American English*. Ph.D. dissertation, Northern Arizona University.

Guy, Gregory R. 1980. Variation in the group and the individual: The case of final stop deletion. In William Labov (ed.), *Locating Language in Time and Space*, 1–36. New York: Academic Press.

Haak, Anu, Evi Juhkam, Marja Kallasmaa, Arnold Kask, Ellen Niit, Piret Norvik, Vilja Oja, Aldi Sepp, J. Simm & Jüri Viikberg. 1989. *Väike murdesõnastik II* [Small Estonian dialect dictionary], edited by Valdek Pall. Tallinn: Valgus. Keele ja Kirjanduse Instituut.

Hansen, Björn. 2005. How to measure areal convergence: A case study of contact-induced grammaticalization in the German–Hungarian–Slavonic contact area. In Björn Hansen

https://doi.org/10.1017/S0332586517000191 Published online by Cambridge University Press

& Petr Karlík (eds.), *Modality in Slavonic Languages: New Perspectives*, 219–237. München: Sagner.

Hansen, Björn. 2014. The syntax of modal polyfunctionality revisited: Evidence from the languages of Europe. In Elisabeth Leiss & Werner Abraham (eds.), *Modes of Modality: Modality, Typology, and Universal Grammar* (Studies in Language Companion Series 149), 89–126. Amsterdam & Philadelphia, PA: John Benjamins.

Hansen, Björn & Ferdinand de Haan (eds.). 2009. *Modals in the Languages of Europe*. Berlin & New York: Mouton de Gruyter.

Harrell, Frank E. Jr. 2001. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer.

Heine, Bernd & Tanja Kuteva. 2005. *Language Contact and Grammatical Change*. Cambridge: Cambridge University Press.

Holvoet, Axel. 2001. *Studies in the Latvian Verb*. Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego.

Holvoet, Axel. 2009. Modals in Baltic. In Hansen & de Haan (eds.), 199–228.

Hothorn, Torsten, Kurt Hornik & Achim Zeileis. 2006. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics* 15(3), 651–674.

Hothorn, Torsten, Peter Bühlmann, Sandrine Dudoit, Annette Molinaro & Mark J. Van Der Laan. 2006. Survival ensembles. *Biostatistics* 7(3), 355–373.

ISK = Auli Hakulinen, Maria Vilkuna, Riitta Korhonen, Vesa Koivisto, Tarja Riitta Heinonen & Irja Alho. 2004. *Iso suomen kielioppi* [Comprehensive grammar of Finnish]. Helsinki: Suomalaisen Kirjallisuuden Seura.

Janda, Laura [A.]. 2008. Transitivity in Russian from a cognitive perspective. In Galina Kustova (ed.), *Dinamičeskie modeli: Slovo. Predloženie. Tekst. Sbornik statej v čest' E. V. Padučevoj*, 970–988. Moscow: Jazyki slavjanskoj kul'tury.

Kaalep, Heiki-Jaan & Kadri Muischnek. 2002. *Eesti kirjakeele sagedussõnastik* [A frequency dictionary of Estonian]. Tartu: Ülikooli Kirjastus.

Kalnača, Andra. 2013. Darbības vārda *vajadzēt* modālā semantika [Modal semantics of verb VAJADZĒT 'to need']. *Vārds un tā pētīšanas aspekti* 17(1), 80–88. Liepāja: Liepājas Universitāte.

Kask, Arnold. 1984. *Eesti murded ja kirjakeel* [Estonian dialects and written language]. Tallinn: Valgus.

Keenan, Edward. 1976. Towards a universal definition of 'subject'. In Charles N. Lee (ed.), *Subject and Topic*, 303–334. New York: Academic Press.

Kehayov, Petar. 2009. Olema-verbi ellipsist eesti kirjakeeles [Ellipsis of the Verb *olema* in Written Estonian]. *Emakeele Seltsi aastaraamat* 2008(54), 107–152.

Kehayov, Petar, Liina Lindström & Ellen Niit. 2011. Imperative in interrogatives in Estonian (Kihnu), Latvian and Livonian. *Linguistica Uralica* 47(2), 81–93.

Kehayov, Petar & Reeli Torn-Leesik. 2009. Modal verbs in Balto-Finnic. In Hansen & de Haan (eds.), 363–401.

Klaas-Lang, Birute & Miina Norvik. 2014. Balti areaali tüpoloogilisi sarnasusi morfosüntaksi valdkonnas [Typological similarities in morphosyntax in the Baltic area]. *Keel ja Kirjandus* 8–9, 590–608.

Klavan, Jane, Marja-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.

https://doi.org/10.1017/S0332586517000191 Published online by Cambridge University Press

Koit, Enn. 1963. Eitus saarte murdes [Negation in insular dialect]. In Paul Ariste & Valdek Pall (eds.), *Nonaginta: Johannes Voldemar Veski 90. sünnipäevaks 27. juunil* 1963. Special issue of *Emakeele Seltsi toimetised 6*, 136–147.

Koptjevskaja-Tamm, Maria & Bernhard Wälchli. 2001. The Circum-Baltic Languages: An Areal-typological Approach. In Dahl & Koptjevskaja-Tamm (eds.), vol. 2, 615–750.

Kortmann, Bernd. 2010. Areal variation in syntax. In Peter Auer & Jürgen E. Schmidt (eds.), *Language and Space: Theories and Methods. An International Handbook of Linguistic Variation*, 837–864. Berlin & New York: Mouton de Gruyter

Kyröläinen, Aki-Juhani. 2013. *Reflexive Space: A Constructionist Model of the Russian Reflexive Marker*. Turku: University of Turku.

Laitinen, Lea. 1992. *Välttämättömyys ja persoona. Suomen murteiden nesessiivisten rakenteiden semantiikkaa ja kielioppia* [Necessity and person: The semantics and grammar of necessive structures in Finnish dialects]. Helsinki: Suomalaisen Kirjallisuuden Seura.

Lindström, Liina. 2017. Partitive subjects in Estonian dialects. In Liina Lindström & Tuomas Huumo (eds.), *Grammar in Use: Approaches to Baltic-Finnic*: Special issue of *Journal of Estonian and Finno-Ugric Linguistics* 8(2), 191–231.

Lindström, Liina, Mervi Kalmus, Anneliis Klaus, Liisi Bakhoff & Karl Pajusalu. 2009. Ainsuse 1. isikule viitamine eesti murretes [The first person singular reference in Estonian dialects]. *Emakeele Seltsi aastaraamat* 2008(54), 159–185.

Lindström, Liina, Maarja-Liisa Pilvik, Mirjam Ruutma & Kristel Uiboaed. 2017. On the use of perfect and pluperfect in Estonian dialects: Frequency and language contacts. In Sofia Björklöf & Santra Jantunen (eds.), *Plurilingual Finnic: Change of Finnic Languages in a Multilinguistic Environment*, 51–89. Helsinki: Finno-Ugrian Society.

Lindström, Liina, Maarja-Liisa Pilvik, Mirjam Ruutma & Kristel Uiboaed. 2015. Mineviku liitaegade kasutusest eesti murretes keelekontaktide valguses [The use of the compound past tenses in Estonian dialects in the light of language contacts]. *Võro Instituudi toimõndusõq* 29, 39–70.

Lindström, Liina & Ilona Tragel. 2010. The possessive perfect construction in Estonian. *Folia Linguistica* 44(2), 371–399.

Lindström, Liina, Kristel Uiboaed, Virve-Anneli Vihman. 2014. Varieerumine *tarvis*/*vaja*-konstruktsioonides keelekontaktide valguses [Variation in Estonian ‘need’-constructions in the light of language contacts]. *Keel ja Kirjandus* 8–9, 609–630.

Lindström, Liina & Virve-Anneli Vihman. 2017. Who needs it? Variation in experiencer marking in Estonian ‘need’-constructions. *Journal of Linguistics* 53(4), 789–822.

Mets, Mari. 2010. *Suhtlusvõrgustikud reaalajas: võru kõnekeele varieerumine kahes Võrumaa külas* [Social networks in real time: Variation in spoken Võro in two villages of Võru County] (Dissertationes philologiae estonicae universitatis tartuensis 25). Tartu: Tartu Ülikooli Kirjastus.

Metslang, Helena. 2013. Coding and behaviour of Estonian subjects. *Journal of Estonian and Finno-Ugric Linguistics* 4(2), 217–293.

Metslang, Helle. 2009. Estonian grammar between Finnic and SAE: Some comparisons. *Sprachtypologie und Universalienforschung* 62(1/2), 49–71.

Metslang, Helle & Liina Lindström. 2017. Essive in Estonian. In Casper de Groot (ed.), *Uralic Essive and the Expression of Impermanent State* (Typological Studies in Language 119), 57–91. Amsterdam & Philadelphia, PA: John Benjamins.

Metsmägi, Iris, Meeli Sedrik & Sven-Erik Soosaar. 2012. *Eesti etümoloogiasõnaraamat* [Estonian etymological dictionary]. Tallinn: Eesti Keele Sihtasutus.

Must, Mari. 1987. *Kirderannikumurre* [North Eastern Coastal Dialect]. Tallinn: Valgus.

Must, Mari & Aili Univere. 2002. *Põhjaeesti keskmurre: häälikulisi ja morfoloogilisi peajooni* [North Estonian Mid Dialect: Phonological and morphological features] (Eesti Keele Instituudi Toimetised 10). Tallinn: Eesti Keele Instituut.

Neetar, Helmi. 1964. Aluse ja öeldise ühildumist mõjutavatest teguritest eesti murretes [Subject and predicate agreement in Estonian dialects]. *Emakeele Seltsi aastaraamat* 1964(10), 151–166.

Neetar, Helmi. 1970. Määrsõnalisest täiendist eesti murretes [Adverbial attribute in Estonian dialects]. *Emakeele Seltsi aastaraamat* 16(1970), 195–206.

Nurkse, Rein. 1937. *Adjektiiv-atribuudi kongruentsist eesti keeles* [Agreement of adjectival attribute in Estonian]. Tartu: Akadeemiline Emakeele Selts.

Pajusalu, Karl, Tiit Hennoste, Ellen Niit, Peeter Päll & Jüri Viikberg. 2009. *Eesti murded ja kohanimed* [Estonian dialects and place names], 2nd edn. Tallinn: Eesti Keele Sihtasutus.

Penjam, Pille. 2006. *tulema*-verbi grammatilised funktsioonid eesti kirjakeeles [Grammatical functions of the verb *tulema* in written Estonian]. *Keel ja Kirjandus* 1, 33–41.

Penjam, Pille. 2011. Eesti kirjakeele subjektilised ja adessiivadverbiaaliga TARVITSEMA-konstruktsioonid [*Tarvitsema*-constructions with subject or adessive adverbial in written Estonian]. *Keel ja Kirjandus* 5, 505–525.

Pilvik, Maarja-Liisa. 2016. *olema* + Vmine konstruktsioonid eesti murretes [On *olema* + *Vmine* constructions in Estonian dialects]. *Keel ja Kirjandus* 6, 429–446.

Pilvik, Maarja-Liisa. 2017. Deverbal *-mine* action nominals in the Estonian Dialect Corpus. In Liina Lindström &Tuomas Huumo (eds.), *Grammar in Use: Approaches to Baltic-Finnic*: Special issue of *Journal of Estonian and Finno-Ugric Linguistics* 8(2), 295–326.

R Development CoreTeam. 2013. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. http://www.R-project.org/.

Ruutma, Mirjam, Aki-Juhani Kyröläinen, Maarja-Liisa Pilvik & Kristel Uiboaed. 2016. Ambipositsioonide morfosüntaktilise varieerumise kirjeldusi kvantitatiivsete profiilide abil [Descriptions of the morphosyntactic variation of ambipositions by means of quantitative profiles]. *Keel ja Kirjandus* 2, 92–113.

Saareste, Andrus. 1938. *Eesti Murdeatlas: Atlas Des Parlers Estoniens* [Estonian dialect atlas]. Tartu: Eesti Kirjanduse Selts.

Saareste, Andrus. 1955. *Petit Atlas Des Parlers Estoniens: Väike Eesti Murdeatlas* [Small Estonian dialect atlas]. Uppsala: Almqvist & Wiksell.

Seržant, Ilja A. 2012. The so-called possessive perfect in North Russian and the Circum-Baltic area: A diachronic and areal account. *Lingua* 122, 356–385.

Seržant, Ilja A. 2015a. The independent partitive as an Eastern Circum-Baltic isogloss. *Journal of Language Contact* 8, 341–418.

Seržant, Ilja A. 2015b. Dative experiencer constructions as a Circum-Baltic isogloss. In Peter Arkadiev, Axel Holvoet & Björn Wiemer (eds.), *Contemporary Approaches to Baltic Linguistics*, 325–348. Berlin & Boston, MA: de Gruyter Mouton.

Seržant, Ilja A. & Valgerður Bjarnadóttir. 2014. Verbalization and non-canonical case marking of some irregular verbs in *-ē- in Baltic and Russian. In Artūras Judžentis, Tatyana Civjan & Maria Zavyalova (eds.), *Balai ir slavai: dvasinių kultūrų sankritos/ Балты и славяне: пересечения духовных культур* [The Balts and Slavs: Intersections of spiritual cultures], 218–242. Vilnius: Versmės.

https://doi.org/10.1017/S0332586517000191 Published online by Cambridge University Press

Strobl, Carolin, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin & Achim Zeileis. 2008. Conditional variable importance for random forests. *BMC Bioinformatics* 9(1), 307.

Strobl, Carolin, Anne-Laure Boulesteix, Achim Zeileis & Torsten Hothorn. 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics* 8(1), 25.

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

Szmrecsanyi, Benedikt. 2013. *Grammatical Variation in British English Dialects: A Study in Corpus-based Dialectometry* (Studies in English Language). Cambridge: Cambridge University Press.

Tagliamonte, Sali A. & R. Harald Baayen. 2012. Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change* 24(2), 135–178.

Torn-Leesik, Reeli & Virve-Anneli Vihman. 2010. The uses of impersonals in spoken Estonian. *SKY Journal of Linguistics* 23, 301–343.

Uiboaed, Kristel. 2010. Statistilised meetodid murdekorpuse ühendverbide tuvastamisel [Statistical methods for phrasal verb detection in Estonian Dialects]. *Eesti Rakenduslingvistika Ühingu Aastaraamat* 6, 307–326.

Uiboaed, Kristel. 2013. *Verbiühendid eesti murretes* [Verb constructions in Estonian dialects]. Tartu: University of Tartu Press.

Uiboaed, Kristel. 2016. *Spatial Visualization with R*. Spatial-visualization-with-r 1.0. https://doi.org/10.5281/zenodo.51473.

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

Vaba, Lembit. 2011. Kuidas läti-eesti keelekontakt on mõjutanud eesti murdekeele grammatikat ja sõnamoodustust [Impact of Latvian–Estonian language contacts of the grammar and word-formation of Estonian dialects]. *Emakeele Seltsi aastaraamat* 2010(56), 204–246.

Van de Velde, Hans & Roeland van Hout. 1998. Dangerous aggregations: A case study of Dutch (n) deletion. In Carole Paradis (ed.), *Papers in Sociolinguistics*, 137–147. Quebec: Nuits Blanches.

Vangsnes, Øystein Alexander. 2007. Scandinavian dialect syntax (before and after) 2005. *Nordlyd* 34(1), 7–24.

Velsker, Eva. 2013. *Aeg* eesti murretes [*Aeg* ‘time’ in Estonian dialects]. *Emakeele Seltsi aastaraamat* (2012)58, 265–295.

Viitso, Tiit-Rein. 2014. Constructions of obligation, duty, and necessity in Livonian. *Journal of Estonian and Finno-Ugric Linguistics* 5(1), 193–214.

Wade, Terence. 2011. *A Comprehensive Russian Grammar*, 3rd edn. Revised and updated by David Gillespie. Oxford: Wiley-Blackwell.

Wälchli, Bernhard. 2011. The Circum-Baltic languages. In Bernd Kortmann & Johann van der Auwera (eds.), *The Languages and Linguistics of Europe: A Comprehensive Guide*, 325–340. Berlin & Boston, MA: de Gruyter Mouton.

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

https://doi.org/10.1017/S0332586517000191 Published online by Cambridge University Press

Wolk, Christoph & Benedikt Szmrecsanyi. 2016. Top–down and bottom–up advances in corpus-based dialectometry. In Marie-Hélène Côté, Remco Knooihuizen & John Nerbonne (eds.), *The Future of Dialects: Selected Papers from Methods in Dialectology XV* (Series: Language Variation 1), 225–243. Berlin: Language Science Press.

https://doi.org/10.1017/S0332586517000191 Published online by Cambridge University Press