Variation of Verbal Constructions in Estonian Dialects

Kristel Uiboaed
University of Tartu
kristel.uiboaed@ut.ee

Liina Lindström
University of Tartu
liina.lindstrom@ut.ee

Cornelius Hasselblatt
University of Groningen
c.t.hasselblatt@rug.nl

Kadri Muischnek
University of Tartu
kadri.muischnek@ut.ee

John Nerbonne
University of Groningen
j.nerbonne@rug.nl

University of Tartu
Department of Estonian and General Linguistics
Jakobi 2
51014 Tartu
Estonia

University of Groningen
Center for Language and Cognition
Oude Kijk in 't Jatstr. 26
Rijksuniversiteit Groningen
Postbus 716
NL 9700 AS Groningen
The Netherlands

Abstract

Traditional Estonian dialect classifications are based on the phonology, morphology, and lexis, and there are very few studies about syntax available. The present paper is the first quantitative syntactic study of Estonian dialects. We concentrate on constructions consisting of finite and non-finite verbs, and we apply contemporary statistical methods to explore the syntactic variation. Our results show that even bare token frequencies can identify syntactic patterns quite well, and that analyses exploiting collostructional methods makes the variational patterns even clearer. We use correspondence analysis and clustering to detect geographic influence on variation. The results suggest a syntax-based classifications of dialects that differs from the traditional classifications based mainly on phonology and lexis. Our data reveals systematic differences between eastern and western dialects at the syntactic level, while analyses based on phonology and lexis distinguish mainly between northern and southern dialects. The western dialects make more use of analytic constructions consisting of a finite and a non-finite verb form.

1. Introduction

The main aim of this paper is to contribute to Estonian dialect syntax and Estonian dialectology. In addition, we add a case to the discussion on how geography may differentially influence the different systemic levels of language, e.g., phonology vs. syntax. Finally we report on the application of statistics developed in corpus linguistics, which play a facilitating role in the analysis.

To date Estonian dialects have mainly been studied from the perspectives of phonology and lexis, and only very few studies about syntax are available. The present study aims to contribute to filling in that gap. The present paper studies syntactic variation in Estonian dialects, more specifically the variation of a special kind of verbal construction: finite and non-finite verb constructions (V_{fin} + Non-Fin).

Non-finite verb forms are regularly formed verb forms which can have a number of different functions in a sentence and which lack many typical verbal traits. Non-finites may be further classified as infinitives, participles, converbs and action nominals (verbal nouns or *masdars*) (Ylikoski 2003). The Estonian language has a variety of non-finite forms and they

often form different constructions with various finite verbs, some of which have undergone grammaticalization (*e.g.* Habicht *et al.* 2010, Tragel 2003, Tragel & Habicht *to appear*). The traditional grammar of Estonian (Erelt *et al.*, 1993) defines V_{fin} + Non-Fin constructions as kind of periphrastic verbal construction, where one component modifies the meaning of the other, and the type of sentence is determined by the whole construction. The finite verb expresses the modality, aspect, causativity or manner of the state of affairs expressed by the non-finite verb (Erelt *et al.* 1993). Similar constructions have received attention in standard Estonian (Habicht *et al.* 2010, Penjam 2008, Tragel 2003, Tragel & Habicht *to appear*, Metslang 2006, Metslang 1993a, among others). The current paper explores this kind of construction and its variation in Estonian dialects; more specifically, we explore which constructions consisting of a finite verb and a non-finite verbal category (*e.g.* Eng. *want to go, let go*) are most common in different dialects. Some finite verbs that occur in these constructions are so strongly grammaticalized that they have acquired auxiliary verb functions when they co-occur with certain non-finite forms.

We concentrate on verbal constructions for several reasons. First, a study of particle verbs in dialects clearly indicated distinct differences between eastern and western dialects where eastern dialects used considerably fewer particle verbs than western dialects (Uiboaed 2010), *i.e.* were less analytic in that respect. This is different than the dialect classifications based on phonology, morphology and lexis, where the biggest differences occur between southern and northern dialects. Clarifying the usage of verbal constructions enables us to get more evidence for these tendencies. Second, just reading corpus texts, which our study relies on, one gets the impression that western part dialects use more V_{fin} + Non-Fin constructions where finite verbs are strongly grammaticalized opposed to eastern dialects where morphological way of expressing seems to be more common. We ask whether this kind of tendency is general or whether only certain constructions or finite verbs are grammaticalized in specific functions.

Third, we wished to use corpora of orthographically transcribed dialect speech both because they more naturally reflect genuine dialect use than e.g. questionnaire data, as Szmrecsanyi & Kortmann (2009) argue, but also because they provide frequency data. As sociolinguistics has shown for decades, variation is often reflected in frequency rather than categorical differences (Labov 1966). Having decided to use corpora as the data on which to base analyses, we need to focus on phenomena that can be extracted automatically and in large numbers. Verbal complementation patterns fit the bill quite nicely.

This leads us to note a further contribution of this study. Dialect syntax has been enjoying a growth in interest of late (Barbiers, *et al*. 2005, Heap 2000), but most of the work has focused on the analysis of large databases of syntactic features that experts have compiled. There has been much less work that has proceeded from corpora (Szmrecsanyi & Kortmann 2009), and present study is innovative in expanding that line of work to include a new language (Estonian) and new sort of syntactic variation, that of collostructions, i.e., affinities between lexemes and particular slots in constructions. We shall be more concrete about the combinations we examine below (Sec. 5).

Our central research question is whether Estonian dialects group syntactically just as they do phonologically, morphologically and lexically. Our hypothesis, following Uiboaed (2010), is that they do not; we expect to find more distinct differences between eastern and western dialects as opposed to the traditional North-South distinction. We assume the differences may arise for instance from the stronger Germanic influence in the West. The reason for the east-west distinction on the basis of the syntax is not clear but we can assume that it may be based on the one hand, on the more conservative nature of the eastern dialects (which have been in contact with eastern Finnic languages, mainly with Votic (Must 1987, Alvre 2000) while western dialects have had more influence from old written Estonian which have had a strong Germanic influence (cf. Alvre 2000). On the other hand, western dialects, especially Insular dialect, have had strong contacts with Swedish. Thus, the overall tendency of preferring analytic verbal constructions in western dialect could be attributed to the influence of Germanic languages which may have come directly or via Old Written Estonian. Additionally, we clarify whether and how different constructions vary in different dialects and how dialects differ in terms of observed constructions and their frequencies. The present paper is only concerned with the categories of non-finite forms, not with the actual verbs used in that forms, so we only describe V_{fin} + Non-Fin (finite form lemmata and non-finite verb class) constructions.

If we are correct that syntactic variation is distributed differently with respect to geography than phonology, lexis and morphology, than the interesting question arises as to why this should be. After all, we expect the diffusion of innovations to proceed along similar lines of dense communication (Bloomfield 1933:Ch.3.4), and therefore also expect the resulting distributions of variation to be similar. This is not a focus of the current paper, but we return to this discussion in the closing section.

From a methodological point of view, the paper compares two methods from corpus linguistics for detecting constructions. First we detect constructional patterns based only on raw normalized frequencies of V_{fin} + Non-Fin combinations, assuming that if these two forms frequently appear together in the same clause they also form a semantic and syntactic unit. Second we apply the collostructional methods developed by Stefanowitsch & Gries (2003). We use Fisher's exact test to gauge collostructional strength between non-finite verb morphological category and finite verb lemma. We describe differences in results when these two methods are applied.

The second section of the paper gives a brief overview of the relevant aspects of the Estonian language and its dialects. The third section describes existing non-finite forms in Estonian language. We then describe our data sources and construction extraction process in fourth section, and the fifth part describes the methods used in the current study. The sixth section presents the results of constructional analysis, and in the seventh part we present some results of the qualitative analysis.

2. Estonian Language and Dialects

Estonian is a Finno-Ugric language belonging to the Finnic branch. The closest relatives to the language are the Livonian and Votic, which are presently nearly extinct. The closest languages to Estonian still used for everyday communication are Finnish, Karelian, and Veps. There are about one million Estonian speakers in the world. Estonian has been influenced strongly by Indo-European languages, so that traits atypical for a Finno-Ugric language can be detected at all the levels of structure. Estonian has a very complex morphological system, which is typical for a Finno-Ugric language (Erelt *et al.* 2000).

The area where Estonian is spoken is rather small, but the differences among the traditional dialects are substantial. There are slightly different classifications of Estonian dialects available, but for the purpose of comparison the present paper proceeds from the most detailed classification, which is the one used in the corpus of Estonian dialects (see the fourth section), according to which (i) the North Estonian dialect group includes Insular, Western, Mid and

Eastern dialects; and (ii) the South Estonian group consists of the Mulgi, Tartu, Seto and Võru[1] dialects. The Northeastern Coastal dialect group is part of the North Estonian group and includes the Coastal and Northeastern dialects. These dialect groups can be divided to more than a hundred sub-dialects. (Lindström & Pajusalu 2003) The map in Figure 1 presents the traditional Estonian dialect areas. Traditional dialect classifications distinguish most significantly between northern and southern dialects, and the biggest differences are in phonology and lexis.

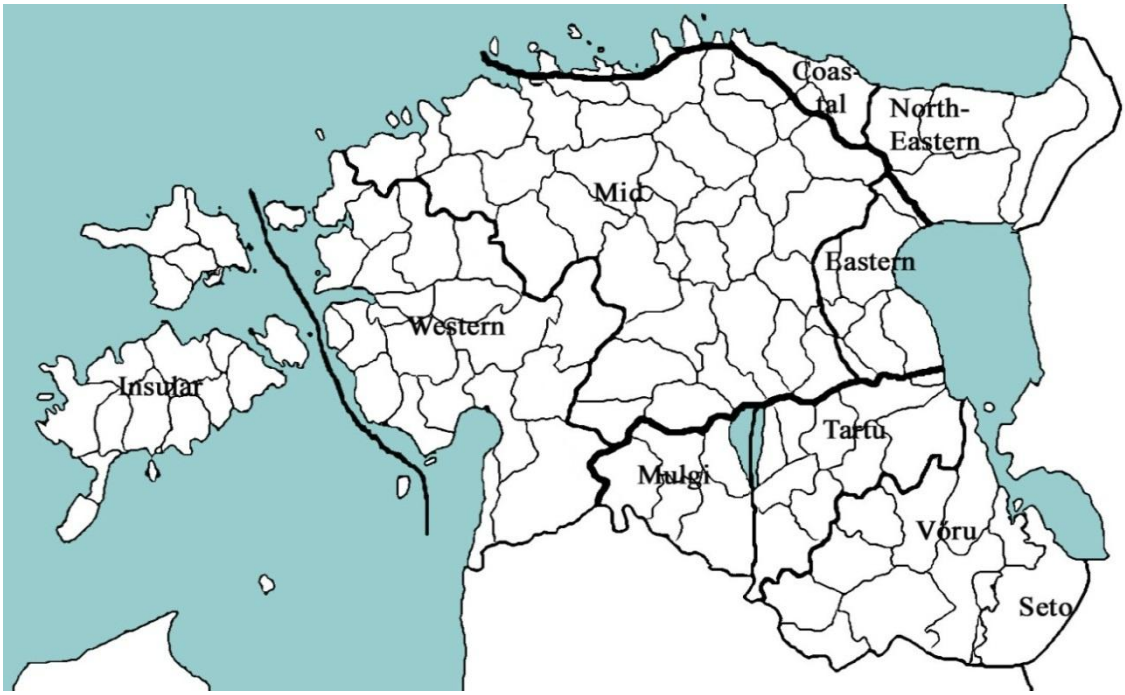

Figure 1. Estonian dialect areas. The North-Estonian dialect group includes the Coastal, Eastern, Insular, Mid, North-Eastern, and western dialects. The South-Estonian group includes the Mulgi, Tartu, Seto, and Võru dialects.

3. Non-finite Verbal Categories in Estonian

Non-finite verbs can form different constructions with finite verbs. These constructions can be either complex predicates or argument structure constructions as in *lähen sööma* 'I go to eat'.

[1] Drawing a distinction between Seto and Võru has been a complicated issue in Estonian dialectology. The main difference between two dialects lies in pronunciation and in
lexis. Pajusalu *et al.*(2009) do not find joining these two dialects acceptable due to remarkable territorial and cultural differences. Seto speakers are Eastern orthodox, as opposed to the mostly protestant Võru speakers. Seto also has stronger Russian influence on vocabulary and pronunciation (Pajusalu *et al*. 2009: 187).

The borderline between these two groups is not an exact one, so that verb + verb constructions rather make up a continuum (Sahkai & Muischnek 2010). Their behaviour in dialects has not been studied so far and the present paper attempts to fill that gap by clarifying the possible constructional patterns in different dialects of Estonian. Table 1 represents all the non-finite forms and their formatives in Estonian.

NON-FINITE FORMS	PERSONAL	IMPERSONAL
Participles		
Present	*v*	*dav tav*
Past	*nud*	*dud tud*
Supine		
Illative (2. infinitive)	*ma*	*dama tama*
Inessive	*mas*	
Elative	*mast*	
Translative	*maks*	
Abessive	*mata*	
Infinitive	*da a ta*	
Gerund	*des es tes*	

Table 1. Non-finite forms and their formatives in Estonian (Viitso 2003)

The following section gives a short overview of some non-finite verb forms in standard Estonian and is based completely on Erelt *et al.* (1993, 2000) and Erelt (2003). Only a brief overview of non-finite verb forms' semantics and syntactic functions is given, as the main goal is to illustrate the non-finite forms in Estonian. We concentrate on the non-finite forms as only these are relevant in our later analysis.[2] The present paper is concerned only with non-finite forms and which finite verbs they co-occur with in a clause. We were interested only in possible combinations and their frequencies, so we attempted to detect variation patterns with respect to these, i.e. measuring which constructions are more common to different dialects and in which constructions vary.

3.1 Participles

In standard Estonian two participles are distinguished: the present and past participle, both of which can have personal and impersonal forms. **Present participles** can occur as attributes and predicatives and are inflected for case and number, i.e. they function similarly to adjectives. **Past**

[2] For a more thorough overview of non-finite forms and their classification, see Ylikoski (2003).

participles[3] can also occur as attributes and predicatives, but in addition they are regularly used to form compound tense forms with finite forms of the verb *olema* ‘to be’ (1a–b).

(1) a. *Eksam* *on* *kirjuta-**tud**.*
exam.NOM be.3SG.PRS write-PPP
‘Exam has been written.’

b. *Ma* *olen* *seda* *eksamit* *kirjuta-**nud**.*
I be.1SG.PRS this.PART exam.PART write-APP
‘I have written that exam.’

*3.2 Supine (*ma-*infinitive, 2. Infinitive, 2INF)*

The 2INF is the traditional headword for verbs in Estonian dictionaries. It usually appears as an adverbial, but it may also take on other syntactic functions. The 2INF expresses relative future or entering into a process and it also occurs in sentences as an adverbial indicating destination (2).

(2) *Ta* *läks* *jaluta-**ma**.*
(s)he go.3SG.PST walk-2INF
‘(S)he went for a walk’

The 2INF forms not only inchoative constructions with a variety of verbs, e.g. ‘to start’, ‘to go’, ‘to come’, ‘to stay’, *etc.* (3a) but also causative constructions with the verbs ‘to put’, ‘to hit’, and ‘to remain’ among others (3b). It also forms the modal verb construction with the verb *pidama* ‘to have to’ (3c).

(3) a. *Me* *hakkasime* *koju* *mine-**ma**.*
we start.1PL.PST home.ILL go-2INF
‘We started to go home.’

b. *Ta* *pani* *tule* *põle-**ma**.*
(S)he put.3SG.PST light.PART burn.2INF
‘(S)he turned on the light (lit. he put the light on)’

c. *Me* *peame* *tööle* *mine-**ma**.*
We must.1PL work.ALL go.2INF
‘We have to go to work.’

[3] From here on we use active and passive past participles and APP and PPP glosses respectively, but traditionally impersonal and personal participles are more common.

The supine form can also take inessive, translative, abessive, and elative case endings. For instance, the inessive form of 2INF can express the progressive meaning (Metslang 1993a) (4).

(4) *Ilmad* *on* *soojene-**ma**-**s**.*
wethers be.3PL.PRS warm up-2INF-INE
'The weather is getting warmer (every day)'

3.3 Infinitive (da-infinitive, 1INF)

The 1INF can serve various syntactic functions in a sentence. It can occur as a subject, object (5a), adverbial, or predicative.

(5) a. *Ma* *oskan* ***laul-da**.*
I can.1SG.PRS sing-1INF
'I can sing.'

Modal verbs typically form the constructions with the 1INF. These modal verbs do not determine the presence, meaning or form of the subject, which is only determined by the semantics of the non-finite verb (Erelt *et al.* 1993; Erelt 2001). The 1INF can form constructions with various modal verbs (*võima* 'can', *tohtima* 'may', *saama* 'get, become', *tulema* 'to come' in modal meaning) (6).

(6) *Ma* *võin* *sind* *aida-**ta**.*
I can you.PART help-1INF
'I can/am able to help you.'

As mentioned above the description given here of the non-finite forms and their functions is far from exhaustive. All these forms can occur in different functions and may form constructions with other verbs. We have more thorough analyses of 1INF and 2INF constructions in standard Estonian (see Penjam 2008), but we are interested in how these constructions are used in dialects.

4. Data

4.1 The Corpus of Estonian Dialects

The number and scope of comparative studies about Estonian dialects have been rather small due to the lack of suitable data sources. In order to improve that situation the University of Tartu and the Institute of Estonian Language in Tallinn started a joint project in 1998 to compile an electronic data source for that purpose. The main aim of the project was to build a large corpus to conduct studies about the phonological and grammatical structure of Estonian dialects supported by electronic data processing (Lindström, Pajusalu 2003). The corpus of Estonian dialects (CED) is an electronic database containing authentic dialect texts from all ten major dialects of the Estonian language.

The CED consists of:

- dialect recordings;
- texts in standard Finno-Ugric phonetic transcription;
- texts in simplified transcriptions;
- morphologically annotated texts in XML-format;
- a database containing information about interviewers and recordings; and
- some texts with syntactic annotation.

The informants in the CED were chosen on the basis of their social properties: they are typically poorly educated elderly people who have themselves lived all their lives in one place, and whose parents have, as well. In older dialectal research, such informants have been seen as ideal for representing older local dialect speech.

CED is a textual record of spoken spontaneous language. Special features of speech have been taken into account, and all discourse particles, word repetitions, pause fillers, corrections *etc.* have been transcribed. The recorded interviews are traditional dialect interviews, where the interviewer interviews the informant on familiar territory (the informant's home or backyard). The oldest recordings date back to 1938, but the majority of the interviews were recorded during the 1960 –70s. Although the older texts have been recorded in the studio, the nature of the interviews is the same compared to later recordings.[4]

[4] 2% of words come from texts recorded in 1938 (five texts all from different dialect areas). These texts were recorded in studio, but the nature of the interviews and topics are exactly the same as in other interviews. The

At the moment when the present study was conducted the CED contained 1.245.000 words of text in phonetic transcription of which 665 000 words had been morphologically annotated. The morphologically annotated CED is freely available on the web: *www.murre.ut.ee* and also via the international dialect syntax webpage Edisyn: *http://www.dialectsyntax.org/index.php/edisyn-othermenu-51/emk.* More detailed information about the CED and principles of tagging can be found in Lindström *et al*. (2009).

4.2 Construction Extraction

Data was obtained from the morphologically annotated CED described above. In order to extract finite verb lemma and non-finite verb morphological category pairs, it was necessary to set clause boundaries because verbs form a construction only if they co-occur in the same clause. For that purpose, the syntactic parser of the Estonian language (Müürisep 2000), adapted for dialect parsing, was applied (Lindström & Müürisep 2009). Candidate data was extracted by forming all the possible combinations of the finite verb lemma with the non-finite verb category within one clause. These pairs do not necessarily occur next to each other as illustrated in (7).

Frequency counts for the analysed data were calculated as follows:

1) Only category (morphological) tags for the non-finite forms were used, ignoring the specific the verb (lexeme) itself.
2) All the occurrences of finite verbs were counted, regardless of their tense, mood, number *etc.* Only the lemmas of finite verbs were used for the analysis.
3) Frequency counts for the whole construction were based on the co-occurrence of the finite and non-finite forms in the same clause.

To calculate the precision of the extraction process we randomly chose 500 words from every dialect, 5000 words altogether. The precision of construction extraction was 80%, but when we removed low frequency combinations (less than 3 occurrences) as we did in our final analysis it rose to 92%. The dialects did not differ greatly in the precision with which constructions were extracted.

Estonian word order shows considerable variability. The finite verb has a tendency to occur in the second position in main clauses and in the final position in some subordinate

---

informants are, just as in later recordings, poorly educated, elderly local people. To get the maximum out of our data we included these texts, as we are convinced that such a small amount of data cannot change the big picture.

clauses; at the same time word order is dependent on information structure (see (7) and (8)) (Lindström 2005; Tael 1988). Different parts of the constructions can be displaced in the clause still carrying the same meaning:

(7) *Ta* ***hakkas*** *kõva*
(s)he start.3SG.PST loud

häälega ***laul-ma***.
voice.COM sing-2INF
'When he started to sing loud.'

(8) *Kui ta* ***laul-ma*** ***hakkas***.
When (s)he sing-2INF start.3SG.PST
'When he started to sing'

To avoid regarding *hakkas laulma* and *laulma hakkas* as different constructions, all the combinations were recorded in a cononcial order based on the non-finite verb form (morphological category). Only the grammatical category of the non-finite verb form and the dictionary form of the finite verb form were taken into account. The final list of constructions for every dialect looks like example (9), where for instance the inchoative construction *laulma hakkama* described above has become the 2INF and *hakkama* 'start, become' construction (*ma hakkama*) among other 2INF and *hakkama* constructions:

(9) inf saama *1INF 'become/get'*
ma pidama *2INF 'have to'*
ma kutsuma *2INF' invite'*
ma tulema *2INF 'come'*
ma hakkama *2INF 'start'*
tud olema *PPP 'be'*
…

The first column is the non-finite verb form information as it is annotated in the CED and the second column is a lemma of the finite verb. Table 1 in Section 3 presented the abbreviations for non-finite forms also used in the CED. The fact that the CED is morphologically annotated enables us to use morphological categories like the ones presented in (9).

Constructions can also be formed from three verbs, but we concentrated on two-verb constructions. It is not a trivial task to identify constructions consisting of three verbs and our method produced two constructions where a three verb construction occurred.

(10) *Ma* ***hakkasin*** ***jooksma*** ***minema***.
I start.1SG.PST run.2INF go.2INF
'I was about to go runnig (lit. I started to go to run)'

The example (10) illustrates a three-verb construction (*hakkasin jooksma minema*) where our method detected two V_{fin} + Non-Fin constructions (*hakkasin minema* 'I started to go' and *hakkasin jooksma* 'I started to run'). However, the problem was not substantial therefore we did not exclude these from our analysis.

5. Methodology

5.1 Collostructional analysis

Additionally to raw frequencies we applied collostructional analysis to extract the constructions. Collostructional methods are family of quantitative corpus linguistic methods developed by Stefanowitsch & Gries (2003). Collostructional methods are similar to more known collocation finding methods (Evert 2005; 2008), which measure the statistical strength between two words. Collostructional methods measure also the strength between two linguistic units, but include syntactic and/or semantic factors. The word 'collostruction' is a blend of 'collocation' and 'construction' (Stefanowitsch & Gries 2003; 2005), and the analytical focus is on the relationship between words and constructions they participate in (Stefanowitsch & Gries 2003). Their analysis adopts the terminology of Construction Grammar (Fried & Östman 2004, Goldberg 1995, Kay & Fillmore 1999) and is normally applied to constructions and the words they occur with. The present study applies the method on a more schematic level, investigating only the relationship between the category of a non-finite verb and the finite verb lemma it co-occurs with.

We chose covarying collexeme analysis, one of a family of collostructional techniques, which measures the statistical strength between a non-finite verbal morphological category and a finite verb lemma.

To clarify which constructional patterns are genuine, and not randomly co-occurring items, the *Coll.analysis 3.2* program developed for collostructional analysis (Gries 2007) was applied to calculate the collostructional strength for each non-finite form and finite verb combination. This program calculates the association strength between two units, in our case the morphological category of the non-finite form and the finite verb lemma, based on their frequencies. We made calculations separately for all ten dialects and chose Fisher's exact test to measure the association strength.

The association strength between non-finite categories and the finite verbs that co-occur in the same clause is calculated based on information of the sort illustrated in Tables 2 and 3. Table 3 illustrates the construction 1INF + *tahtma* 'want' in the eastern dialect. This combination occurs in the eastern dialect five times (there were five clauses containing this combination). There are six occurrences of the verb *want* with other non-finite forms. An infinitive occurs 47 times with a finite verb other than *want,* and there are 1096 combinations of non-finite form plus finite verb involving neither 1INF or the verb *want*.

	Finite verb Y in slot 2	**All other finite verbs in slot 2**
Non-finite form X in slot 1	Freq (X slot1 + Y slot2)	Freq (X slot1 + ⌐Y slot2)
All other non-finite forms in slot 1	Freq (⌐X slot1 + Y slot2)	Freq (⌐X slot1 + ⌐Y slot2)

Table 2. Two-way contingency table for co-varying collexeme analysis

	finite verb *tahtma* 'want'	**other finite verbs**	**row totals**
1INF	5	47	52
other non-finite forms	6	1049	1055
column totals	11	1096	N=1107

Table 3. Two-way contingency table for 1INF + *tahtma* 'want' construction in the Eastern dialect.

Measures of association strength compare the frequency with which two items co-occur to the frequency with which they might be expected to co-occur based on chance. We calculate the probability of the elements occurring by chance under the assumption that the two elements are statistically independent. All the information that is needed for computation is contained in the tables 2 and 3. Since the 1INF category occurs 52 times in 1107 clause, we estimate its frequency as 52/1107, or about 5%; for the form *tahtma* we estimate its frequency as 11/1107, or

about 1%. If these two sorts of elements were statistically independent, we would expect to see them in combination about a relative frequency of approx. 0.05 × 0.01 = 0.0005. Wiechmann (2008) compares over 40 measures of association strength for use in corpus linguistics, and Fisher's Exact Test (FET) performs extremely well.

We therefore use FET to gauge the collostructional strength between a non-finite category and a finite verb; the higher the value, the stronger the relation between two units. We use a 95% confidence interval to determine the threshold of the FET value we require, and combinations that do not reach this threshold are considered too weakly associated and therefore excluded from the analysis. We emphasize that the present study does not compare FET values to each other. As we used FET to measure the association strength and different dialects have different amount of material in the corpus, FET values were not comparable to each other. We used FET only to detect genuine constructions, assuming that combinations that have lower values are not genuinely interdependent. We then compared the normalized frequencies of constructions that had a FET value above the threshold we set.

5.2 Correspondence analysis

To detect similar groups of dialects and to identify their distinctive features we applied Correspondence Analysis (CA). CA is a method of data analysis that attempts to describe tabular categorical data and presents a multi-dimensional dataset in a two-dimensional plot; it is often used to analyse frequency tables. CA attempts to find latent patterns in regular frequency tables by calculating distances separately between rows and between columns and presenting the results in a two-dimensional space. Although CA in principle enables the researcher to use more than two dimensions, it is rare that more are ever used. The stronger the association between two data points is, the closer they appear on a CA map. (Cichocki 2006; Greenacre 2007; Lebart, Salem & Berry 1998). Axes do not have any frequency interpretation on the CA map; instead they only present two dimensions of the multidimensional dataset and percentages that show the inertia (comparable to variance) explained by these two dimensions. One should only detect patterns on the CA map that the data supports. We applied the method to illustrate the similarities and differences between dialects based on the non-finite and finite verb constructions and their frequencies in ten dialects. Closeness of dialects on the CA map indicates the strong association (similarity) between these dialects in terms of constructions and their frequencies. If two dialects

use similar constructions and also these constructions have similar frequencies these two dialects appear close on the CA map. This enables us to see which groups dialects form and to determine whether they are similar to the traditional dialect classification or, alternatively, whether there are any differences. Dialects are interpreted as similar if the same constructions appear in them, and constructions are interpreted as similar when they tend to appear in the same dialects.

5.3 Strength of Signal

To measure the consistency of the frequency table and the strength of the geographical signal in the data, Cronbach's Alpha was calculated (Cronbach 1951). Cronbach's Alpha is a consistency measure that shows whether the number of analysed items is sufficient for getting consistent results. Its value ranges from 0 to 1 – the higher the value more reliability the dataset is. The generally accepted threshold is 0.7.

Local incoherence was calculated to measure the lack of coherence in data. The smaller the measure, the more coherent dataset is (Nerbonne & Kleiweg 2007). For calculating Cronbach's Alpha and Local incoherence a *Gabmap* software package developed at the University of Groningen was used (Nerbonne *et al.* 2011).

6. Analysis of the Constructions Observed in the Data

This section presents an analysis of the data in two different ways. First, we will give an overview of the analysis based only on the normalized text frequencies of finite verb lemma and non-finite verb form combinations, assuming that if these two forms co-occur in the same clause, they form a construction. The second analysis takes into account the results of the collostructional analysis with FET as the measure of association strength.

6.1 Analysis of constructions based on the normalized frequencies

Conducting the analyses based only on normalized V_{fin} + Non-Fin pair frequencies was encouraging as the quality measures Cronbach's Alpha 0.85 and Local incoherence 0.16 were promising.

To explore the differences between different dialects in terms of finite lemma and non-finite verb constructions, frequency counts for different combinations were extracted

automatically from the corpus. All the combinations with raw frequency less than three were excluded from the analysis. In order to make frequencies in different dialects comparable – as there are different amounts of material available for different dialects in CED – some normalization was needed. Therefore all the frequencies were normalized based on the average corpus size (61 312 words). For instance, the construction 2INF + *minema* 'to go' occurred 39 times in the eastern dialect. The size of the whole eastern part of the CED was 43 965 running words. After normalization there were, for example, 39 occurrences of 2INF + *minema* in eastern dialect, resulting in the normalized frequency of 54 = 39 x 61312/43965. After removing low frequency (<3) combinations, the list contained 120 different types of potential constructions.

An advantage of this approach is that it enables us to include more potential constructions in analysis; a disadvantage that it also includes a lot of noise. This noise is mainly produced by the fact that the parser does not set clause boundaries perfectly. The accuracy of the parser is quite good: only 0,4% of clause boundaries were mis-detected (Lindström & Müürisep 2009; Müürisep 2011, personal communication), but the accuracy of the parser is also dependent on the dialect and the nature of the text. Sometimes finite and non-finite pairs crossing clause boundaries are mis-detected, which results in the inclusion of non-constructions. The problem is more serious with frequent verbs and frequent non-finite forms, for instance passive and active past participles. Defining a clause in a spoken language is not an easy task, as there are lot of repetitions and corrections, all of which can cause the overdetection of constructions.

Figure 2 illustrates the results of CA applied to this data table. The dialects are presented in capitals and coloured green and the constructions are in lower case. The further the items are from each other in the scatterplot, the more different they are. The *x*- and *y*-axes show proportions of inertia (explained variance) explained by the first two dimensions. The South-Estonian dialect group (Mulgi, Tartu, Seto, Võru) shows considerably more variation than the northern one. The *y*-axis dimension suggests one group containing Võru, Tartu, Eastern, Northeastern and Seto dialects and the other one consisting of Mid, Insular, Coastal, Western and Mulgi dialects. The *x*-axis dimension distinguishes Mid, Insular, Coastal, Western, Northeastern, Seto and then Eastern, Võru, Tartu, Mulgi dialects. Two of the groups are clearly visible: a lower left quadrant consisting of the Mid, Western, Insular and Coastal dialect and an upper right quadrant containing the Eastern, Tartu and Võru dialects. Mulgi, Seto and Northeastern dialects do not form natural classes.

There seems to be a big difference between Mulgi and Seto dialects; both also differ from Tartu and Võru, but in different ways. The difference between Seto and Võru is surprising as they are considered to be the same in most dialect classifications (Pajusalu *et al.* 2009).

One has to keep in mind that the interpretation of distances between the sites and constructional items is not as straightforward as comparing the sites and constructions separately. The CA graphs sites and constructions separately and just superimposes the one graph on the other. The proximity of sites and constructions items is an approximation. For instance, the constructions *ma_heitma* (2INF + 'to bed down') and *inf_jõudma* (1INF + 'to manage') are more characteristic of Võru, Tartu and Eastern dialects.

6.2 Analysis Based on the Normalized Frequencies and FET Values

The second analysis takes into account the association strength scores, namely the p-values from Fisher's exact test (FET), which are regarded as indications of the constructional strength between a non-finite verb's morphological category and a finite verb lemma.

The procedure for that analysis begins just as the previous ones: all the two category combinations were generated. We experimented with three measures: Odds-ratio, Fisher's exact test (FET) and additionally Minimum Sensitivity (Pedersen, Bruce 1996; Wiechmann 2008). Finally FET was chosen because it performed well on all ten dialects, because it is especially suitable for working with language data (Pedersen 1996), and because it has been applied and found to be suitable in constructional studies (Stefanowitsch & Gries 2003; Gries & Stefanowitsch 2004). FET values were calculated based on the raw frequencies and normalization was done after the extraction of constructions with the high association score.

FET was applied as follows:

1) Separately for all 10 dialects, we calculated FET p-values for all the non-finite verb form and finite verb lemma combinations. As a result, we got 10 different lists of constructions ordered according to their collostructional strength, i.e FET values computed by *Coll.analysis 3.2.* program (Gries 2007).
2) We set the threshold to the collostructional strength on the significance level of $p < .05$. Combinations that did not fulfil this criterion were excluded.
3) From here on we did not use FET values in our analysis anymore. We analysed only the constructions and their normalized frequencies, i.e. the frequencies of the

constructions that satisfied the requirements of a significance level of $p < .05$ and raw frequency greater than 3.

4) At this point we have (normalized) frequency tables of different constructions for every dialect (low frequency combinations and less significant combinations both removed).
5) We compare only the constructions and their frequencies in different dialects, e.g. what are low- and high-frequency constructions in different dialects. Which constructions are present in some dialects and not in others?

We took the FET p-value as an indicator, because it helps to remove some noise from the data, e.g. high raw-frequency combinations consisting of high frequency verbs and forms. It provides also more evidence for claiming that certain combinations are constructions and others are not. The final list included the 57 different types of constructions

The advantage of this approach is that it reduces the noise in the data, but it also excludes some potential constructions, i.e., those that fail to reach the threshold for association scores. Association measures take into account category and verb frequencies separately, which is definitely a considerable advantage over not using association strength. Cronbach's Alpha for this dataset was 0.75 and Local incoherence 0.15.

Again correspondence analysis was conducted and Figure 3 presents the results. Here the differences within the southern group remain. But interestingly clear clusters form between the eastern and western dialects along the second dimension (y-axis). The total inertia also increases with this analysis, which indicates a stronger relationship between sites and constructions. It is remarkable that northern and southern dialects do not form clear clusters. East and North-East dialects, traditionally classified as belonging to the northern group, seem to be closer to southern dialects in their constructional nature.

Just as we saw in the first analysis, bare frequencies yield some geographical patterns on the basis of the constructional variation, but the results are not very clear due to the amount of noise in the data. Including association strength measures reduces the noise in the data and makes geographical patterns more visible. So we may conclude that using only bare frequencies gives us a lot of information about constructional variation, but incorporating association measures definitely clarifies these tendencies further (although at the cost of some loss of information as we shall see in Section 7).

6.3 Clustering

To examine the differences between eastern and western dialects from another perspective, the clustering techniques available in the *Gabmap* software package were applied (Nerbonne *et al*. 2011). The analysis aims to explore the eastern and western differences further. Both dataset preparations – bare normalized frequencies and filtered by FET – were analysed. The aim of the analysis was to test whether clustering also recognizes different eastern and western dialect groups.

We applied a fuzzy clustering method (Nerbonne *et al.* 2008), which adds various amounts of random noise to the distance matrix as it re-clusters. The probabilistic dendrograms in Figures 4 and 5 illustrate the results. Clusters that appear many times are particularly stable ones. The percentages on the dendrogram show how many times clusters appeared during the noisy iteration process. We may be confident of clusters that have been detected 100 times (100%), while clusters detected infrequently may be artefacts of the analysis.

The two dendrograms present quite similar results. Dendrograms clearly distinguish eastern and western dialects: Insular, Mid and Western dialects are included in one cluster and rest of the dialects in another. The division within the eastern dialects is not clear, but it still provides interesting results. In the first dendrogram (taking account the FET values) the Eastern, Võru and Tartu dialects quite clearly form a cluster. The Eastern dialect is traditionally included with the northern dialect group, which should be more similar to Mid, Insular and Western dialects than to the South-Estonian group. The second dendrogram (using only bare frequencies) groups together Eastern, Seto, Võru and Coastal dialects quite strongly together. Similar results were provided by all the other clustering techniques available in *Gabmap*.

Both clustering results associate Eastern, Coastal and North-Eastern dialects more strongly with the southern dialect group, i.e. the dialects form clear East-West groups, which is different from the traditional dialect classifications where the stronger distinction is made between the North and South.

7. Differences between Eastern and Western Dialects

Our two analyses confirm that there are strong differences between the eastern and western dialects. The present section explores these differences in more detail. In order to analyse the differences between East and West we divided dialects into two groups, where the eastern dialects included Eastern, Mulgi, Northeastern, Seto, Tartu and Võru dialects (based on a total of almost 237,700 words), and the western group consisted of the Coastal, Insular, Mid and Western dialects (based on a total a bit over 375,000 words). The classification was made based on both CA and fuzzy clustering analyses results. As the results were not the same in all analyses, we also took into account previous dialect classifications and a previous study on a similar topic (Uiboaed 2010). We relied more on the results of CA and previous classifications and did not make final judgments based solely on the clustering results as they were not always clear. For instance, in the Figures 4 and 5 the Coastal dialect is included in the Eastern group, but CA results (Figures 2 and 3) clearly place the Coastal dialect in the same group as the Insular, Mid and Western dialects. As clustering results do not give a very stable signal about the Coastal dialects, we decided to include the coastal dialects in the Western group based on the very clear CA results. We formed these two groups based only on our statistical analyses, and we shall now continue with qualitative linguistic analyses to explore which constructions are responsible for the division proposed by statistical analyses.

The following section briefly presents some constructions that distinguished the eastern and western dialects especially well. If some non-finite forms do not appear in the list, it means that there are no considerable differences between eastern and western dialects, or that there were too few cases for drawing conclusions. We present only some of the meanings of these constructions and do not consider our survey or our discussion to be exhaustive. The same finite and non-finite pair may have several meanings or functions, but we only present the most common ones found in our data. The exploration of the linguistic nuances of each construction must remain a goal for future work. Conclusions are drawn based on both the FET-based and the normalized frequency analyses. We also point out some differences between these two analyses.

7.1 General differences[5]

The biggest difference was the overall frequency of construction usage; western dialects use considerably more V_{fin} + Non-Fin constructions. Raw frequency analysis turned up 8542 occurrences of constructions from the western and 5758 from the eastern dialects (FET analysis 6559 and 5056 respectively).[6] This also confirms the result of Uiboaed (2010), which revealed that eastern dialects used fewer particle verbs, i.e. were less periphrastic in that respect. Our results indicate the same; western dialects tend to use analytic expressions more (than the eastern varieties); the frequencies of single constructions are also mostly higher in the western group. The exploration of linguistic nuances of each construction is future work.

7.2 Constructional Differences

The tables 4 and 5 present finite verb and non-finite morphological category pairs and some central meanings that these forms can carry. We present the constructions that are more common in both groups. Categorization under eastern or western does not mean that this construction can not appear in the other group. We present only some constructions and their central meanings that were more common to the western group and others more common to the eastern group. The first column presents the form (non-finite morphological category and finite verb occurring with that); the second column presents some central meanings of these constructions and we also present some examples which are indicated in the brackets with meanings. Numbers in the brackets refer to examples presenting this type of use of the construction.

[5] To facilitate reading examples are standardized and transcription symbols have been removed from here on, as these symbols carry only the pronunciation information, which is not relevant here. Every example includes the notation whether it belongs to the eastern or the western group, e.g. W-MID, E-SET. Abbreviations are presented in the appendix.

[6] These are normalized frequencies of two groups. Note that these number differ from those presented in previous sections due to the different bases of normalization.

WESTERN	**some central meanings**
1INF + *tohtima* 'can, may'	modality (11)
1INF + *võima* 'can'	modality (12)
1INF + *tahtma* 'want'	intention, wish, modality (13)
1INF + *laskma* 'let'	enabling-obligation (Penjam 2008) (14), causative (Kasik 2001)
1INF + *olema* 'be'	passive, impersonal, modality, semi-fixed mental verb constructions (15)
2INF + *juhtuma* 'to happen'	non-volitionality, unintentionality (16)
2INF + *hakkama* 'to start'	inchoative (17), future
2INF + ajama 'to lead, to drive'	causative (Kasik 2001) (18)
2INF + *saama* 'to get, become'	resultative (19), succeeding, fixed expression with the verb *hakkama* 'to start' in meaning 'to cope'
2INF + *panema* 'to put'	causative (20)
2INF inessive + *käima* 'to go'	habitual (21)
PPP + *saama* 'to get, become'	passive (22), impersonal, resultative, possessive perfect
PPP + *olema* 'to be'	passive (23), impersonal, resultative, possessive perfect (Lindström & Tragel 2010)

Table 4. Constructions more common in the western group of dialects

(11) *tema* *ligi* *ei* ***tohi*** ***minna*** *lapsed* (W-MID)
(s)he.GEN close NEG can.3PL.NEG go.1INF children
'children cannot go near him/her'

(12) *aga* *siis* ***vöis*** ***laolda*** (W-INS)
but then can.SG.PST sing.1INF
'but then one could sing'

(13) *sie* ***tahab*** *uold* ***saada*** (W-COA)
this want.3SG.PRS care.PRT get.1INF
'this needs (lit. wants) to be taken care of'

(14) *ja* *mind* *ei* ***lastagi*** *neile*
and I.PART NEG let.1INF them.ALL

süija ***viija*** (W-MID)
eat.1INF bring.1INF
'and they do not let me to get them food'

(15) *aga* *nüid* *enam* *seda*
but now anymore this.PART

vene *kielt* ***ei*** ***ole***

Russian.GEN language.PART NEG be.3SG.NEG.PRS

kuulda *nüd* (W-MID)
hear.1INF now
'but one cannot hear Russian (here) anymore'

(16) *ma* ***juhtusin*** *vahel* *natukene*
I happen.1.SG.PST sometimes a bit

iljemaks ***jääma*** (W-WES)
later stay.2INF
'sometimes I happened to be late'

(17) *tema* ***akkab*** *kohe* ***nutma*** (W-COA)
(s)he start.3SG.PRS soon cry.2INF
'(s)he is about to start crying'

(18) ***aas*** ***külmetama*** *inimese* (W-COA)
lead.3SG.PST freeze.2INF person.GEN
'it made (lit. led) the person to freeze

(19) *siis* ***said*** *pulmalesed* *koeu* ***menema*** (W-MID)
then get.3PL.PST wedding guests home go.2INF
'then wedding guests were able to start going home'

(20) *vahõl* *mõnõ* *õhta* ***pandi***
Sometimes some evening put.IMPS.PST

tüe ***seismä*** (W-INS)
work.GEN stand.2INF
'sometimes, some evening the work was stopped (lit. was put to stand)'

(21) *vanad* *inimest* ***käisitte*** *Suomes*
old.NOM people go.3.PL.PST Finland.SG.INE

kala ***püüdamas*** (W-COA)
fish.SG.PART catch.2INF.INE
'old people went fishing to Finland'

(22) *mõisnikkud* ***saavad*** *ära* ***kaotud*** (W-WES)
estate owners.NOM become.3.PL.PRS away lose-PPP
'estate owners are being lost'

(23) *juussed* ***ollid*** *sedamodi* ***leigetud*** (W-INS)
hair be.3PL.PST this way cut.PPP
'hair was cut this way'

EASTERN	**some central meanings**
1INF + *jõudma* ‘to reach, to manage’	physical and mental ability (24)
2INF + *minema* 'to go'	inchoative (25)
2INF + *tulema* 'to come'	modality, motion (26)
2INF_abessive + *jääma* 'to leave, to remain'	negative passive (27)

Table 5. Constructions more common in the eastern group of dialects

(24) *kess* *soo* *unõnäo* *mullõ* *ärr*
who this.GEN dream.GEN me.ALL away

jõvass ***juudustada*** (E-SET)
manage.3SG.PRS.COND tell.1INF
‘who would manage to explain (lit. tell) me that dream’

(25) *nä* ***lätsivä*** *sinnä* *rahha*
they go.3PL.PST there money.PART

tiinmä ***kauplõmma*** (E-SET)
earn.2INF trade.2INF
‘they went there to earn money and to trade’

(26) *vanamiis* ***tull*** *hää meelegä* ***mängmä*** (E-VOR)
old man.NOM come.3SG.PST with pleasure play.2INF
‘old man came to play with pleasure’

(27) *mu* *silm* ***jäi*** ***nägemädä*** (E-VOR)
my eye remain.3SG.PST see.2INF.ABE
‘I didn’t see (that) / my eyes remained without seeing (that)’

As mentioned above, this list is not exhaustive, and there is still more to discover about these constructions’ meanings. The following discussion points out some differences between two analyses – the normalized frequency analysis and the FET-filtered analysis.

The modal verb *saama* ‘get, become’ and 1INF construction was more frequent in the western group when we count only its frequency. When we filtered the collostructions by requiring a definite strength of association, its importance rose slightly in the eastern group, which means that the construction was not extracted in some of the western dialects, so that it appears to be more strongly associated in the eastern group. Qualitative analysis showed that the construction is present in all dialects and is quite common in both groups. It is a very polysemous construction usually carrying impersonal, passive and modal meanings. It can also form some

fixed expressions and some semi-fixed expressions with certain mental verbs (*to hear, to see*, *to feel*).

1INF + *olema* ‘to be’ in (15) was not detected in any of dialects when we applied collostructional analysis. Qualitative analysis showed that the construction exists in both dialect groups and that it is more common in the western one.

2INF + *saama* ‘get, become’ (19) was another construction not detected by the association strength measure. Qualitative analysis showed that the construction is present in all dialects, but has a very low frequency. It still seems to be more common in the western group.

2INF in the inessive case and *olema* ‘to be’ can carry progressive or proximative meaning (Erelt & Metslang 2009; Metslang 1993a; Metslang 1993b), and the construction is more common in the eastern group but was not detected at all on the basis of collostructional analyses. The same applies to the 2INF in abessive case and *jääma* ‘to stay, remain’ in (27).

As we see from the tables 4 and 5 there are only a few constructions more common to the eastern group. That means that this group uses morphological means to express the same meanings. Alternatively they may turn to light verb constructions (Muischnek 2006), which were beyond the scope of that study.

The reason why syntactic variation divides mainly along an eastern vs. western dividing line (instead of the traditional line dividing the North and South) remains unclear thus far. We can assume that it may be based, on the one hand, on the more conservative nature of the eastern dialects which have been in contact with eastern Finnic languages, mainly with Votic (Alvre 2000, Must 1987), and on the other, on the more malleable tendencies in the western dialects which have be eninfluenced by written old Estonian. Estonian in its earlier stages was written mainly by German clergymen, and thus, has been influenced strongly by German (cf. Alvre 2000). In the same line, western dialects, especially Insular dialect, have also had strong contacts with Swedish. Thus, it is reasonable to assume that the overall tendency of preferring analytic verbal constructions in western dialect could be explained with the influence of Germanic languages which may have come directly or via Old Written Estonian.

8. Conclusion

The current paper is the first comprehensive quantitative study of Estonian dialect syntax focusing on the variation of finite and non-finite verb constructions. We conducted the corpus

study to explore the syntactic differences among ten Estonian dialects. We first automatically extracted potential constructions from the corpus by examining combinations with particular finite verbs and their non-finite verbal complements, recognized by verbal category. We were only interested in the non-finite verb's category, marked morphologically (as infinitive, participle, *etc.*) but not in the lexical identity of the non-finite verb. The morphologically annotated corpus made extracting this kind of information fairly easy. We achieved a precision of 80%, and after removing low frequency (>3) combinations 92% for this extraction process.

We conducted two analyses. First we only considered the normalized frequencies of the combinations of non-finite category and finite verb lexeme which co-occurred in the same clause. We assumed that if these two co-occur in the same clause they thereby constitute (an instance of) the construction. In the second analysis we first performed Fisher's exact test to calculate the association strength between the lexically specific finite verb and the non-finite category it governs. Association measures also take into account the frequencies of each potential construction and the frequencies of all non-finite categories and their finite governing verbs separately. This provides more evidence for claiming that some combinations are genuine constructions, eliminating others due to the lack of statistical evidence. Our results revealed that just using bare frequencies gives quite reliable information about the constructions and their geographical variation. Including association measures considerably reduced the noise in the data, which is, however, accompanied by a certain amount of loss of information. In some dialects quite frequent constructions did not meet statistical standards when we examined association strength.

In order to detect geographical patterns of dialects based on their constructional nature we applied Correspondence Analysis (CA) and clustering techniques. CA results indicated some distinct differences between eastern and western dialects which clearly differ from the traditional dialect classifications based mostly on phonology, morphology and lexis, where the biggest difference is drawn between the North and South. The western group consisting of Mid, Coastal, Insular and Western dialects was clearly distinguished from the eastern group containing Võru, Tartu and Eastern dialects. North-Eastern and especially Seto and Mulgi varied more within the eastern group. Seto and Mulgi differ considerably from the eastern group, but in a different way. Traditionally these dialects are included in the same group with Võru and Tartu dialects. Seto and Võru are often even considered to be the same dialect, but our syntactic data does not

confirm that claim. Clustering results were similar to CA results. So our analyses reveal that a syntactic perspective can lead to totally different classification of dialects compared to those based on phonology and lexis.

The radically different division is puzzling if we imagine that the diffusion of linguistic features proceeds along similar paths for all features, regardless of their systemic status (phonology, morphology, lexis, syntaxis). Spruit *et al.* (2009) reported fairly similar distributions of phonological, lexical and syntactic variation in the Netherlands ($0.5 < r < 0.65$). On the other hand, several researchers have speculates that there may be different rates of change in syntactic variables as opposed to phonological ones, and these could and indeed should depress the degree to which they would overlap (Dunn *et al.* 2008; Longobardi & Guardino 2009).

To analyse the constructions in more detail we formed two groups: an eastern group including Seto, Tartu, Võru, Mulgi, the Northeastern and the Eastern dialects and a western group consisting of the Insular, Mid, Western and Coastal dialects. Our hypothesis that the western dialects use more verb constructions was confirmed. The construction frequencies were considerably higher in the western group, which likely means that the eastern dialects use more simple tense forms and morphological means to express same meaning. However, it is not completely clear which means eastern dialects use in order to express the same meanings and exploring that remains for the future research. There were very few verbal collostructions which showed slightly higher frequency in the eastern group. The western group uses considerably more periphrastic tense constructions, also inchoative and passive constructions. We can assume that the eastern group uses simple tense forms and morphological way of expressing or totally different constructions instead. It is also the case that the same constructions potentially have different meanings in different dialects. Whether there are any borders between dialects when we include the semantics of each construction and which kind of different meanings constructions carry in different dialects are interesting questions to be answered in the future research.

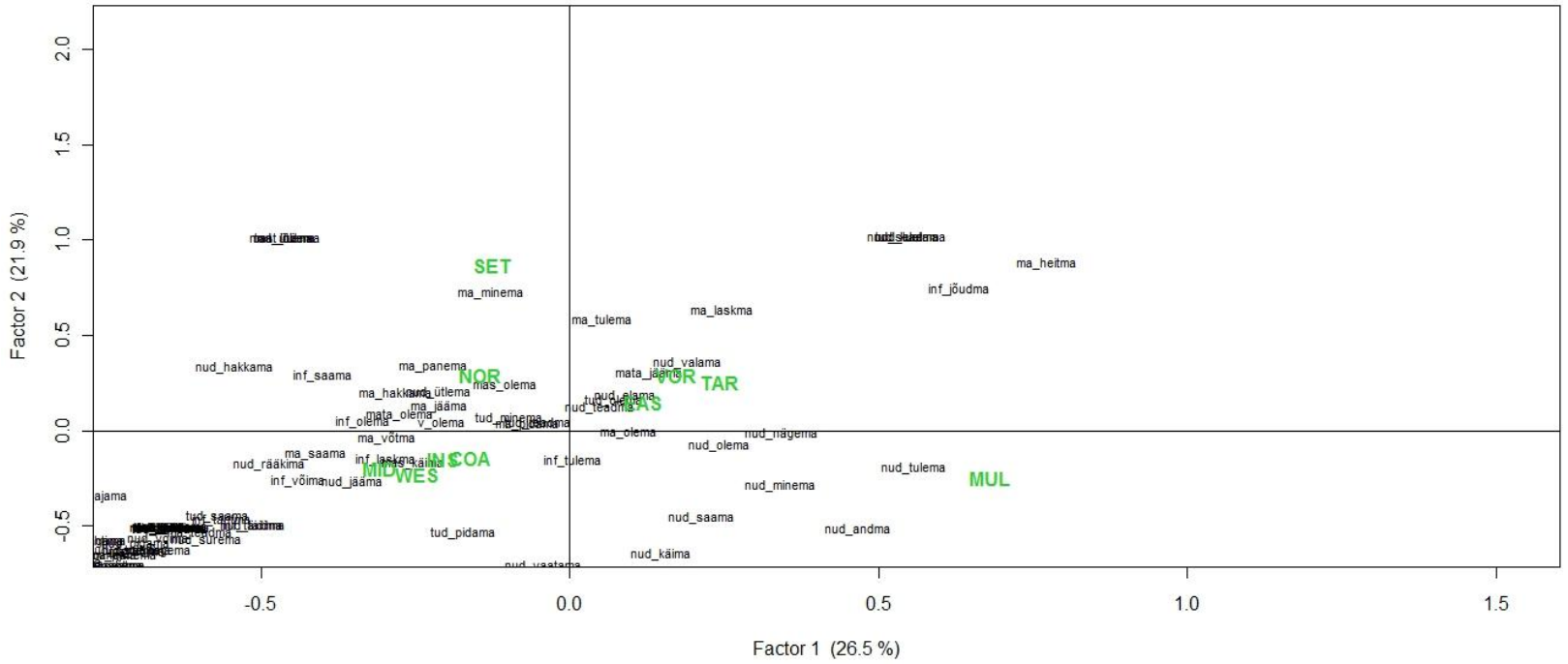

Figure 2. CA plot for constructions in different dialects based on normalized frequencies. Dialect codes: COA – Coastal, EAS – Eastern, INS – Insular, MID – Mid, MUL – Mulgi, NOR – Northeastern, SET – Seto, TAR – Tartu, WES – Western, VÕR - Võru

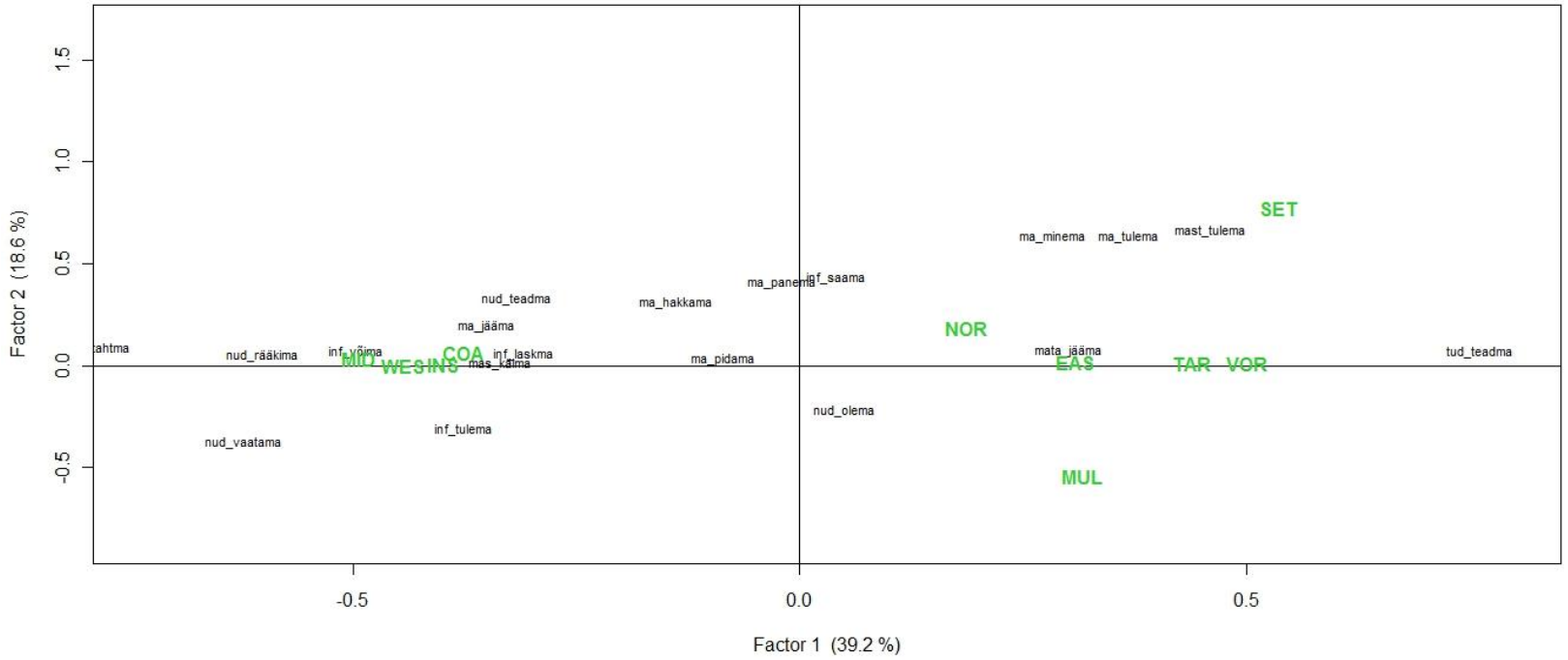

Figure 3. CA plot for constructions in different dialects based on normalized frequencies and FET values. Dialect codes: COA – Coastal, EAS – Eastern, INS – Insular, MID – Mid, MUL – Mulgi, NOR – Northeastern, SET – Seto, TAR – Tartu, WES – Western, VÕR - Võru

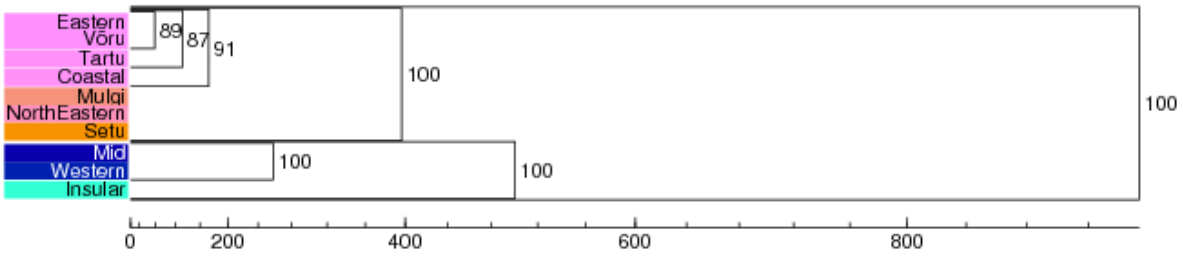

Figure 4. A probabilistic dendrogram clustering dialects based on constructional differences, where only normalized frequency values are considered.

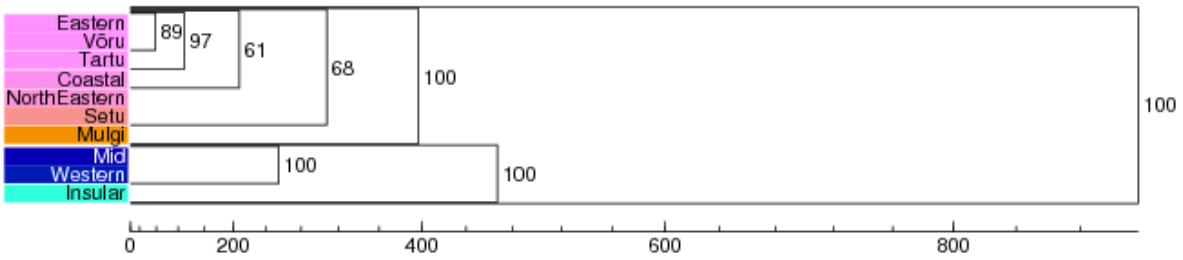

Figure 5. A probabilistic dendrogram clustering dialects based on constructional differences, where FET values were required to be significant.

Abbreviations

Dialect codes

COA	Coastal	SET	Seto
EAS	Eastern	TAR	Tartu
INS	Insular	WES	Western
MID	Mid	VÕR	Võru
MUL	Mulgi	W	western group of dialects
NOR	Northeastern	E	eastern group of dialects

Glosses

1INF	1. infinitive (*da*-infinitive)	INE	inessive
2INF	2. infinitive (*ma*-infinitive, supine)	NEG	negation
ABE	abessive	NOM	nominative
ALL	allative	PL	plural
COM	comitative	PRS	present tense
COND	conditional	PRT	partitive
ELA	elative	PST	past tense
GEN	genitive	SG	singular
ILL	illative	TR	translative

Funding

This work was supported by Estonian Science Foundation [grant 7464], Estonian Ministry of Education and Research [grant SF0180078s08], and European Social Fund's Doctoral Studies and Internationalisation Programme DoRa.

References

Alvre, P. (2000). Kirderannikumurde ja vadja keele ühisjooni. – Inter dialectos nominaque. Pühendusteos Mari Mustale 11. novembril 2000. In. J. Viikberg (ed), (= Eesti Keele Instituudi toimetised 7.). Tallinn: Eesti Keele Sihtasutus, pp. 1–13.

Barbiers, S., Auwera, J. van der, Bennis, H., Boef, E., Vogelaer, G. de, Ham and M. van der. (2005). *Syntactic Atlas of the Dutch Dialects* (SAND), Vol. 1. Amsterdam: Amsterdam University Press.

Cichocki, W. (2006). Geographic Variation in Acadian French /r/: What Can Correspondence Analysis Contribute toward Explanation? *Literary and Linguistic Computing: Journal of the Association for Literary and Linguistic Computing,* 21(4): 529-541.

Cronbach, L.J. (1951). Coeficient alpha and the internal structure of tests, *Psychometrika*, 16: 297-334.

Dunn, M., Levinson, S., Lindström, E., Reesink, G., and Terrill, A. (2008). Structural Phylogeny in Historical Linguistics: Methodological Explorations Applied in Island Melanesia, *Language* 84(4): 710-759.

Erelt, M. (2003). *Estonian language.* Estonian Academy Publishers: Tallinn.

Erelt, M. (2001). Some notes on the grammaticalization of the verb *pidama* in Estonian. *Estonian: typological studies V.* M. Erelt (ed). Publications of the Department of Estonian of the University of Tartu, Tartu, pp. 7-25.

Erelt, M., Erelt, T. and Ross, K. (2000). *Eesti keele käsiraamat,* 2nd edn, Eesti Keele Sihtasutus: Tallinn.

Erelt, M., Kasik, M., Metslang, H., Rajandi, H., Ross, K., Saari, H., Tael, K., and Vare, S. (1993). Eesti keele grammatika II. Süntaks. Tallinn: Eesti Teaduste Akadeemia Keele ja Kirjanduse Instituut.

Erelt, M. and Metslang, H. (2009). Some Notes on Proximative and Avertive in Estonian. *Linguistica Uralica,* vol. 45(3): 178-191.

Evert, S. (2008). Corpora and collocations. In M. Kytö & A. Lüdeling (eds). *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter, pp. 1212-1249.

Evert, S. (2005). *The Statistics of Word Cooccurrences Word Pairs and Collocations*. http://www.bsz-bw.de/cgi-bin/xvms.cgi?SWB12046165 (accessed 25 January 2011).

Fried, M. and Östman, J.-O. (2004). Construction Grammar: A Thumbnail Sketch. In: M.Fried and J.-O.Östman (eds.), *Construction Grammar in a Cross-Language Perspective*. Constructional Approaches to Language 2. Amsterdam: Benjamins, pp. 11–86.

Goldberg, A.E. (1995). *Constructions : a construction grammar approach to argument structure.* University of Chicago press: Chicago.

Greenacre, M. (2007). *Correspondence analysis in practice,* 2nd edn, Capman & Hall/CRC.

Gries, S.T. (2007). *Coll.analysis 3.2. A program for R for Windows 2.x*.

Gries, S.T. and Stefanowitsch, A. (2004). Extending collostructional analysis: A corpus-based perspective on 'alternations', *International Journal of Corpus Linguistics,* 9(1): 97-129.

Habicht, K., Penjam, P. and Tragel, I. (2010). Kas *tahtma* tahab abiverbiks? *Eesti ja soome-ugri keeleteaduse ajakiri,* 2: 115-146.

Heap, D. (2000). *La variation grammaticale en géolinguistique: les pronoms sujet en roman central*. (Lincom Studies in Romance Languages 11) Munich: Lincom Europa Verlag.

Heeringa, W.J. (2004). *Measuring dialect pronunciation differences using Levenshtein distance*. Groningen Dissertations in Linguistics 46.

Kasik, R. (2001). Analytic causatives in Estonian. In M. Erelt (ed), *Estonian: Typological studies V,* Publications of the Department of Estonian of the University of Tartu, Tartu, pp. 77-122.

Kay, P. & Fillmore, C.J. (1999). Grammatical Constructions and Linguistic Generalizations: The What's X Doing Y? Construction. *Language: Journal of the Linguistic Society of America*, 75(1): 1–33.

Labov, W. [1(]1966, 2006). *The Social Stratification of English in New York City*. Cambridge: Cambridge University Press.

Lebart, L., Salem, A. and Berry, L. (1998). *Exploring textual data.* Kluwer Academic Publishers, Dordrecht.

Lindström, L. (2005). *Finiitverbi asend lauses. Sõnajärg ja seda mõjutavad tegurid suulises eesti keeles*. Tartu Ülikooli Kirjastus: Tartu.

Lindström, L. and Müürisep, K. (2009). *Parsing Corpus of Estonian Dialects, Proceedings of the NODALIDA 2009 workshop Constraint Grammar and robust parsing.* Northern European Association for Language Technology, 14.05.2009, pp. 22-29.

Lindström, L. and Pajusalu, K. (2003). Corpus of Estonian Dialects and the Estonian Vowel System, *Linguistica Uralic*, 4: 241-257.

Lindström, L. and Tragel, I. (2010). The possessive perfect ponstruction in Estonian, *Folia Linguistica*, 44(2): 371-400.

Lindström, L., Velsker, E., Niit, E. and Pajusalu, K. (2009). *Mees* 'man', *aeg* 'time' and other frequent words in the Corpus of Estonian Dialects, In M. Kallasmaa & V. Oja. (eds.). *Kodukeel ja keele kodu /Home language and the home of a language*, Tallinn: Eesti Keele Sihtasutus, pp. 91 - 129.

Longobardi, G. and Guardiano, C. (2009). Evidence for syntax as a signal of historical relatedness. *Lingua* 119 (11): 1679–1706. Spec. issue *The Forests behind the Trees*, J. Nerbonne & F. Manni (eds.) .

Metslang, H. (2006). Predikaat ajastut kogemas, *Keel ja Kirjandus* 9: 714-727.

Metslang, H. (1993a). Kas eesti keeles on olemas progressiiv? *Keel ja Kirjandus* 26; (6): 326-334.

Metslang, H. (1993b). Verbitarind ajatähendust väljendamas, *Virittäjä: Journal of Kotikielen Seura* 97(2): 203-221.

Muischnek, K. (2006). *Verbi ja noomeni püsiühendid eesti keeles.* Tartu Ülikooli Kirjastus: Tartu.

Must, M., (1987). *Kirderannikumurre*. Eesti NSV Teaduste Akadeemia, Eesti Keele Instituut. Tallinn: Valgus.

Müürisep, K. (2000). Eesti keele arvutigrammatika: süntaks. Dissertationes Mathematicae Universitatis Tartuensis 22. Tartu Ülikooli Kirjastus: Tartu.

Nerbonne, J., Colen, R., Gooskens, C., Kleiweg, P. & Leinonen, T. (2011). *Gabmap* — A Web Application for Dialectology. Dialectologia, Spec. Issue. 65-89.

Nerbonne, J. & Kleiweg, P. (2007). Toward a Dialectological Yardstick, *Journal of Quantitative Linguistics* 14(2): 148-167.

Nerbonne, J. Kleiweg, P., Heeringa, W., & Manni, F. (2008). Projecting Dialect Differences to Geography: Bootstrap Clustering vs. Noisy Clustering. In: Ch. Preisach, L. Schmidt-Thieme, H. Burkhardt & R. Decker (eds.), *Data Analysis, Machine Learning, and Applications. Proc. of the 31st Annual Meeting of the German Classification Society* Berlin: Springer. pp.647-654.

Pajusalu, K., Hennoste, T., Niit, E., Päll, P., Viikberg, J., Kalvik, **M. and Moorlat, M.** (2009). *Eesti murded ja kohanimed,* 2. edn. Eesti ja üldkeeleteaduse instituut & Eesti Keele Instituut, Eesti Keele Sihtasutus: Tallinn.

Pedersen, T. (1996). Fishing for Exactness. *Proceedings of the South Central SAS User's Group (SCSUG-96) Conference*, pp. 188-200.

Pedersen, T. and Bruce, R. (1996). What to Infer from a Description, *Technical Report 96-CSE-04*.

Penjam, P. (2008). *Eesti kirjakeele da- ja ma-infinitiiviga konstruktsioonid*. Tartu Ülikooli Kirjastus: Tartu.

Sahkai, H. and Muischnek, K. (2010). Liitpredikaadid leksikoni-grammatika kontiinuumil, *Eesti ja soome-ugri keeleteaduse ajakiri ESUKA / Journal of Estonian and Finno-Ugric Linguistics JEF* 1(2): 295 - 316.

Spruit, M. R. Heeringa, W. & Nerbonne, J. (2009). Associations among Linguistic Levels. *Lingua* 119 (11): 1624-1642. Spec. issue *The Forests behind the Trees*, J. Nerbonne & F. Manni (eds.).

Stefanowitsch, A. and Gries, S.T. (2005). Covarying collexemes, *Corpus Linguistics and Linguistic Theory* 1(1): 1-43.

Stefanowitsch, A. and Gries, S.T. (2003). Collostructions: Investigating the interaction of words and constructions, *International Journal of Corpus Linguistics* 8(2): 209-243.

Szmrecsanyi, B. and Kortmann, B. (2009). The morphosyntax of varieties of English worldwide: a quantitative perspective, *Lingua* 119(11):1643-1663.

Tael, K. (1988). Sõnajärjemallid eesti keeles (võrrelduna soome keelega): Läänemeresoome keeleteaduse sümpoosion, Turku, 30.08-2.09. 1988, *Preprint / Eesti NSV Teaduste Akadeemia Ühiskonnateaduste osakond,* Eesti NSV Teaduste Akadeemia Keele ja Kirjanduse Instituut: Tallinn.

Tragel, I. (2003). *Eesti keele tuumverbid.* Tartu Ülikooli Kirjastus: Tartu.

Tragel, I. and Habicht, K. (to appear in *Linguistics*). Grammaticalization of the Estonian saama ‘get’.

Uiboaed, K. (2010). Ühendverbid eesti murretes. *Keel ja Kirjandus* 1: 17-36.

Viitso, T. (2003). Structure of the Estonian language. Phonology, morphology and word formation. *Estonian Language,* Erelt, M (ed) *Linguistica Uralica*, pp. 9-92.

Wiechmann, D. (2008). On the computation of collostruction strength: Testing measures of association as expressions of lexical bias, *Corpus Linguistics & Linguistic Theory* 4(2): 253-290.

Ylikoski, J. (2003). Defining Non-Finites: Action Nominals, Converbs and Infinitives, *SKY Journal of Linguistics* 16:185-237.