Multi-word verbs in a flective language: the case of Estonian

Heiki-Jaan Kaalep
Dept of General Linguistics
University of Tartu
Tartu, Estonia
Heiki-Jaan.Kaalep@ut.ee

Kadri Muischnek
Dept of General Linguistics
University of Tartu
Tartu, Estonia
Kadri.Muischnek@ut.ee

Abstract

This paper describes automatic treatment of multi-word expressions in a morphologically complex flective language – Estonian. It focuses on a special type of multi-word expressions – the verbal multi-word expressions that can function as predicates. Authors describe two language resources – a database of verbal multi-word expressions and a corpus where these items have been annotated manually. The analysis of the annotated corpus demonstrates that the Estonian verbal multi-word expressions alternate in several grammatical categories. Different types of the verbal multi-word expressions (opaque and transparent idioms, support verb constructions and collocations) behave differently in the corpus with regard to the freedom of alternation. The paper describes main types of these alternations and the methods for dealing with them automatically.

1 Introduction

This paper deals with verbal multiword expressions (VMWE) in real texts of a highly inflectional language – Estonian. The main emphasis is on the morphological and syntactic variability of such constructions with some implications and recommendations for their automatic treatment. Once we have a lexicon of VMWEs, large enough to be used in real-life applications (to help with morphological disambiguation, syntactic analysis, machine translation etc.), we need to devise algorithms to actually use them. This in turn requires knowledge about the behavior of VMWEs in real texts.

Estonian language belongs to the Finnic group of the Finno-Ugric language family. Typologically Estonian is an agglutinating language but more fusional and analytic than the languages belonging to the northern branch of the Finnic languages. The word order is relatively free. One can find a detailed description of the grammatical system of Estonian in (Erelt 2003).

In this paper we will focus on a special type of Estonian multi-word expressions, namely those that can function as a predicate in a clause.

This paper is organized as follows. In section 2 we give a brief overview of the VMWEs in Estonian. Section 3 describes the database of the VMWEs and the corpus, where the VMWEs have been manually annotated. Here we will also present the statistics of the VMWEs in the corpus. In section 4 we discuss the variability of these expressions as registered in our corpus and the consequences of these variations for the automatic treatment of the VMWEs. And finally we will make our conclusions in section 5.

2 Types of verbal multi-word expressions in Estonian

A VMWE consists of a verb and 1) a particle or 2) a nominal phrase (usually, but not always, consisting of one noun) in more or less frozen inflectional form, or 3) a non-finite form of a verb. This last combination – verb plus a non-finite verb – remains outside the scope of this paper.

The first combination results in a particle verb. The particle can express location or direction (1), perfectivity (2) etc.

(1) Ta **kukkus** katuselt
alla
S/he fell roof-ABL
down(particle)
'S/he fell off the roof'

(2) Ta **sõi** kõik kommid
ära.
S/he ate all sweets
away(particle)
'S/he ate up all the sweets'

Particle verbs can be either idiomatic as in (3) or non-idiomatic as in (1-2).

(3) Mida nüüd **ette**
võtta?
What now ahead(particle)
take-INF
'What to do now?'

The combinations of a verb and a nominal phrase can be divided into three groups depending on how the components form the meaning of the expression: 1) idiomatic expressions; 2) support verb constructions; 3) collocations.

Idiomatic expressions are usually defined as word combinations, the meaning of which is not the sum or combination of the meanings of its parts. It is meaningful to distinguish between opaque (e.g. English idiom *kick the bucket*) and transparent idioms (e.g. English *pull strings*) as they allow different degrees of internal variability.

Support verb constructions, sometimes also called light verb constructions, are combinations of a verb and its object or, rarely, some other argument, where the nominal component denotes an action of some kind and the verb is semantically empty in this context, e.g. English *make a speech, take a walk.*

The collocations are the fuzziest category. They can be described as VMWEs that do not fit in the previous categories, but still, for some reason, have often been included in dictionaries or are statistically significant combinations of a verb and its argument(s) in the corpus.

In all three groups the non-verbal component is a nominal phrase (not a particle); it can formally be either the object of the verb as in (4), or some other argument as in (5).

(4) Ta **saab** luuletusest hästi
aru
S/he gets poem-EL well
sense-PART
'S/he understands the poem
well.'

(5) Talle **jäävad** luuletused
hästi **meelde**
S/he-ALL remain poems
well mind-ILL
'S/he remembers poems well.'

3 The database and corpus

3.1 Database of VMWEs

Prior to the corpus tagging experiment, a database of Estonian VMWEs (DB) had been compiled, with the aim of creating a comprehensive resource of VMWEs, consisting of 12,200 entries. First, it contained VMWEs from six human-created dictionaries: the Explanatory Dictionary of Estonian (EKSS, 1988-2000), Index of the Thesaurus of Estonian (Saareste, 1979), a list of particle verbs (Hasselblatt, 1990), Dictionary of Phrases (Õim, 1991), Dictionary of Synonyms (Õim, 1993) and the Filosoft thesaurus (http://www.filosoft.ee/thes_et/). In addition, the database had been enriched with VMWEs, extracted semi-automatically from corpora totaling 20 million words, and missing from any of the aforementioned human-made dictionaries. This collocation extraction experiment is described in (Kaalep, Muischnek 2003).

3.2 Corpus

We have a corpus where all the VMWEs have been tagged (by hand). Table 1 shows the composition of the corpus and the number of VMWE instances, compared with the number of sentences and simplex verb instances.

	tokens	sentences	VMWEs	simplex verbs
fiction	104200	9000	3800	21200
press	111100	9500	2400	18000
popular science	99000	7300	1900	15500
total	314300	25800	8100	54700

Table 1. Corpus with VMWEs tagged.

The fiction texts are 2000-word excerpts from Estonian authors from 1980ies. The press files represent various Estonian newspapers (nation-wide and local, dailies and weeklies, quality and tabloid press) from 1995-1999. Popular science comes from the journal „Horisont“, from 1996-2003.

Before tagging the VMWEs, the corpus had been morphologically analyzed and manually disambiguated (Kaalep, Muischnek 2005), making it possible to pre-process the text automatically by tagging the candidate VMWEs in the texts, according to what VMWEs were present in a database of VMWEs. It was then the task of a human annotator to select the right VMWEs, and occasionally to tag new VMWEs, missing from the database and thus having not been tagged automatically. The tagged version was checked by another person, in order to minimize accidental mistakes.

Table 1 shows that the amount and proportion of VMWEs depends on the text class.

Table 2 serves to compare the lexicon of VMWEs based on the corpus with the entries of the DB (the VMWEs from the corpus have been converted to the base form they have in the DB).

A	DB entries	12200
B	A, found in the corpus	2300
C	*hapax legomena* of B	1200
D	new VMWEs	1100
E	*hapax legomena* of D	900

Table 2. VMWEs in the DB and corpus.

First, from rows A, B and D we see that the intersection of the DB and the corpus lexicon is surprisingly small.

The small proportion of VMWEs of the DB that can be found in real texts (compare row B with row A) may be first explained by the small size of the corpus. The second reason is that the human-oriented dictionaries that were used when building the DB implicitly aimed at showing the phraseological richness of the language and thus contained a lot of idiomatic expressions well known to be rare in real-life texts.

The fact that so many VMWEs were missing from the DB was a surprise (compare row D with row A), because, as mentioned earlier, the DB had been enriched with VMWEs from real texts in order to be comprehensive. At the moment, it is not clear what the exact reason is.

The size of *hapax legomena* of new VMWEs also deserves some explanation (compare rows B and C versus D and E).

From the literature, one may find a number of MWU or collocation extraction experiments from a corpus that show that the extraction method yields many items, missing from the available pre-compiled lexicons. Some of the items may be false hits, but the authors (whose aim has been to present good extraction methods) tend to claim that a large number of those should be added to the lexicon.

(Evert 2005) lists a number of authors, who have found that lexical resources (machine readable or paper dictionaries, including terminological resources) are not suitable for serving as a gold standard for the set of MWUs (for a given language or domain). According to (Evert 2005), manual annotation of MWUs in a corpus would be more trustworthy, if one wants to compare the findings of a human (the gold standard) with those of a collocation extraction algorithm.

In lexicography, we may find a slightly conflicting view: not everything found in real texts deserves to be included in a dictionary. Producing a text is a creative process, sometimes resulting in *ad hoc* neologisms and MWUs that are never picked up and re-used after the final full stop of the text they were born in.

Unfortunately these two conflicting views mean that there is no general, simple solution for the problem of finding a gold standard for automatic treatment (extraction or tagging) of MWUs. It is normal that there is a discrepancy between a stand-alone lexicon and the vocabulary of a text.

We believe that the surprisingly high proportion of *hapax legomena* in the set of newly found VMWEs manifests this normal discrepancy of a precompiled lexicon and a text corpus, in our case.

4 Behavior of the VMWEs in the corpus and the problems of their automatic analysis

4.1 Particle verbs

There are two main problems encountered in the automatic identification of the particle verbs. First, as shown in (6-7), the order of the components may vary, and the verb and the particle need not be adjacent to each other, behaving much like particle verbs in German. This varying order and disjuncture of the components is actually characteristic for all the Estonian VMWEs in the text.

(6)Ma **vaatan** need paberid homseks **üle**.
I look these papers tomorrow-TR over(particle).
‘I shall look over those papers by tomorrow.’

(7)Kui sa need paberid **üle** **vaatad**, siis on kõik valmis
If you these papers over(particle) look then is everything ready
‘Once you have looked over those papers, we will be done.’

The second main problem is that most of the particles are homonymous with pre- or postpositions (Estonian has both of them), creating a disambiguation problem, similar to the one concerning the English word *over* in the following examples.

(8) He looked over the papers in less than 10 minutes.
(9) He looked over the fence and saw his neighbor.

Just like in English examples the word-forms *look* and *over* form a phrasal verb *look over* in example (8), but don’t belong together in the same way in example (9), the Estonian verb *vaatama* ‘to look’ and adverb *üle* ‘over’ form a particle verb in the examples (6) and (7), but not in the following example, where *üle* is a preposition:

(10) Ta vaatas **üle aia** ja nägi oma naabrit.
s/he looked over fence-GEN and saw own neighbor-PART
‘S/he looked over the fence and saw his/her neighbor.’

As a pre- or postposition has to be adjacent to the noun phrase that is the constituent of the adpositional phrase, they are usually easier to detect. In (11), however, the invariable word *üle*, that can function both as a particle and a preposition, is positioned before the noun *jõu* ‘force’ in genitive, as if *üle* were a preposition in prepositional phrase *üle jõu* ‘exceeding capabilities’. Actually, it functions as a particle in this clause, forming a particle verb *läks üle* ‘went over’.

(11) Meelitustelt **läks** ta **üle** jõu kasutamisele.
Flattery-PL-ABL went s/he over force-GEN utilization-ALL
‘S/he switched from flattery to violence’

Many of these invariable words that can function either as particles or as pre- and postpositions are quite frequent in the texts. The most frequent simplex verbs are also the most frequent verbal components, forming various VMWEs. The sentences of the written language tend to consist of several clauses. All this results in sentences like (12), where the possible components of particle verbs are scattered across several clauses. In this sentence there are four possible candidate particle verbs: *üle jääma* ‘to have no choice but, lit. remain over’, *üle tegema* ‘to redo, lit. do over’, *ära jääma* ‘be canceled, lit. remain away’, *ära tegema* ‘to accomplish, lit. do away’

(12) Tal ei **jää** muud **üle**, kui töö ise **ära teha**.
S/he-ALL not remain else over(particle)than work-GEN self away(particle) do-INF
‘S/he has no choice but to accomplish the work by her/himself.’

Our preprocessor took only sentence boundaries into account and that resulted in serious overgeneration of possible particle verbs. After experimental tagging of clause boundaries in the texts, the precision of pre-processor improved from 40% to 74% while tagging the particle verbs.

For other types of VMWEs the clause boundaries detection is not so essential. The nominal components of opaque idioms are not so frequent. Some transparent idioms, all support verb constructions and collocations can stretch across clause boundaries, like in (13).

(13) **Kõne**, mille president **pidas**, on mõjutanud meie välispoliitikat.
   Speech that-GEN president held is influenced our foreign-policy-PART
 'The speech held by the president has influenced our foreign policy.'

4.2 VMWEs consisting of a verb and a nominal component

In section 2 we differentiated between three types of VMWEs consisting of a verb and a nominal component, namely idioms, support verb constructions and collocations. All these constructions show considerable variability in the manually annotated corpus. Differently from English, there are no special restrictions on the morphological or syntactic behavior of the verb that is part of an idiom. A VP-idiom, for example the opaque idiom *jalga laskma* 'to run off, lit. to shoot the foot' combines freely with all the morphological categories relevant for the verb, including person, number, tense, mood, non-finite forms and (impersonal) passive. (The latter differs from the English passive - it can be formed from all verbs, having a possible human agent.) The other types of VMWEs – support verb constructions and collocations – have also no restrictions with respect to the verbal inflection.

In this section we will concentrate on the variability of the nominal components of VMWEs – their case and number alternations as registered in the corpus. The case alternation is relevant only for the nominal components that are syntactically in the object position. Our interest in case alternation is motivated by observation that multiword units generally and cross-linguistically tend to be frozen in form. The less variability there is in form, the easier the computational treatment is. We may also draw an analogy between simplex words and multiword units as items in a lexicon. For an inflectional language, every word has an inflectional paradigm, and words with similar paradigms form an inflectional type or class. Variability of VMWEs can be analyzed from the same viewpoint.

From these three types of VMWEs the variation of idioms has received most attention in the literature. Idioms have been regarded as units that can not be given a compositional analysis (e.g. Katz 1973 among others). This view has been opposed later (e.g. Nunberg et. al. 1994). Riehemann (2001) has pointed out that English idioms show considerable variability in text corpora. Describing the automatic treatment of multiword expressions in Basque, Alegria et.al. (2004) show that the support verb constructions in Basque can have significant morphosyntactic variability, including modification of the noun and case alternation. Similar phenomenon (number and case alternation) in Turkish is described in (Oflazer et. al. 2004).

In the following subsections we will briefly describe the phenomenon of the case alternation of the object in Estonian and then discuss the variation of the nominal component of idioms and support verb constructions. Then we will describe the number alternations of the nominal components.

4.3 The case alternation of the object in Estonian

A VMWE often consists of a verb and a noun phrase that is its object syntactically. A few words should be said about the case alternation of the object in Estonian in general (cf also Erelt 2003: 96-97). Three case forms are possible for the object – in singular the object can be either in nominative, genitive or partitive; in plural it can be either in nominative or in partitive. Often the nominative and genitive forms are grouped together under the label ‚total object'.

Partitive is the unmarked form of the object. The partial object, as it is often called, alternates with the total object only in the affirmative clause. In the negative clause only partial object can be used. In the affirmative clause the total object is used only if it denotes definite quantity (is quantitatively bounded) and the clause expresses perfective activity. So, in Estonian, the case alternation of the object is used to express the aspect of the clause – total object can be used if the action described in the clause is perfective:

(14)Mees ehitas **suvilat**
Man built summer-house-PART
'The man built a summer-house/did some summer-house-building.' (imperfective activity)

(15) Mees ehitas **suvila**
Man built summer-house-GEN
'The man built a summer-house.' (perfective activity)

In idioms and support verb constructions the nominal component is only formally or syntactically the object of the verb, semantically it is a part of the predicate. So, it would not be surprising, if such objects wouldn't undergo the case alternations characteristic of the object and would be frozen into the partitive as the unmarked case for the object. Indeed – that is true for the opaque idioms. But for transparent idioms and support verb constructions this is not the case – our corpus data shows that their nominal components can alternate between the forms of total and partial object.

Ca 25% of the transparent idioms in our corpus have their nominal components in the case of the total object:

(16) Esinemisele **pani punkti** ilutulestik.
Show-ALL put full-stop-GEN firework
'The fireworks put an end to the show.'

In the previous example (16) the transparent idiom with the nominal component in the form of the total object was used to describe a perfective action. But the transparent idioms do not form a homogenous group with respect to the case alternation of the nominal component. Some of them behave like regular verb-object combinations; others show irregular variation; and the nominal components of many of them are frozen in the partitive case.

In support verb constructions the case alternation of the object is regularly used to express the aspect of the clause, although the noun denoting an action is non-referential.

(17) Žürii alles **teeb otsust**.
Jury still makes decision-PART
'The jury is still making the decision.' (imperfective)

(18) Žürii **tegi** lõpuks **otsuse**.
Jury made at-last decision-GEN
'The jury made the decision at last.' (perfective)

Some support verb constructions are generally used to refer to the imperfective aspect, to emphasize the process of the action (atelic action), not its result. Such expressions are e.g. *tööd tegema* 'to work, lit. do work-PART' or *sõda pidama* 'fight a war, lit. hold a war-PART'. But, while the nominal component is modified with an appropriate attribute, it can also be in the case of the total object and the support verb expression as a whole then refers to a perfective event:

(19) X ja Y **pidasid** viimase omavahelise **sõja** 17. sajandil.
X and Y held last-GEN mutual-GEN war-GEN 17. century-ADE
'X and Y fought the last war in the 17th century.'

4.4 Number alternations of the nominal components of VMWEs

The nominal component of an opaque idiom in the corpus was always in the same number (singular or plural) as its base form in the DB. For the transparent idioms, the picture was clearly different. Although the nominal component of many transparent idioms does not alternate between singular and plural, there are exceptions, and 14% of the nominal components in the object position and 4% in some other position were in plural.

Support verb constructions, in turn, make extensive use of the number alternations of the nominal component, whereas the plural form of the noun denoting an action can really refer to several events as in (20)

```
(20) **Otsuseid tehti**
konsensuse põhimõttel.
     Decision-PL.PART made
consensus-GEN principle-ADE
     'Decisions were made by
consensus.'
```

4.5 The conclusions for the automatic analysis of VMWEs

The conclusions of the corpus findings for the automatic detection of the VMWEs are the following:
1) The free word order requires that, while detecting automatically the particle verbs in a text, we should be limited with a clause as possible context for the co-occurrences. Using the whole sentence as the possible context would create too much noise, so the detection of clause boundaries is a must.
2) We can treat opaque idioms much like the particle verbs – multi-word units consisting of an inflecting verb and a frozen nominal component that don't cross the clause boundaries.
3) Transparent idioms in the database have to be divided into those enabling their nominal component to appear in the cases of the total object and those, which nominal component is always in partitive. But can the annotator rely on her/his intuition while making such decisions? Rather not, but carrying out corpus research separately on each item is a time-consuming task. It could be a better solution for the transparent idioms to generate all the case forms possible for the object, as the nouns that are part of the idioms are not as frequent as the non-inflecting words that may be particles as well as pre- and postpositions.
4) The nominal component of the support verb constructions can under certain circumstances always be in the form of total object. The nouns denoting action in support verb constructions can also be pluralized. So the best solution for them is to generate all forms of the object, both in singular and in plural, in the database.

5 Conclusion

In this paper we have investigated a subtype of multiword expressions, namely verbal multi-word expressions in a flective language – Estonian.

We have described two linguistic resources – a database of VMWEs and a corpus that has been manually annotated for VMWEs.

These expressions exhibit considerable variation in the corpus. The verb of a VMWE can combine with all the grammatical categories relevant for the verb. The nominal component of a VMWE can alternate in number and case. However, the nominal components of the different types of VMWEs (opaque and transparent idioms, support verb constructions and collocations) have different degrees of freedom.

For a morphologically rich flective language, like Estonian, previous morphological analysis and disambiguation prior to the detecting of the multi-word units in a text is essential.

Credits

The research described in this paper was partly supported by the Estonian Science Foundation (grant 5787).

References

Algeria, Inaki, Olatz Ansa, Xabier Artola, Nerea Ezeiza, Koldo Gojenola, Ruben Urizar. 2004. Representation and Treatment of Multiword Expressions in Basque. *Second ACL Workshop on Multiword Expressions: Integrated Processing*: 48-55.

EKSS 1988 – 2000, *Eesti kirjakeele seletussõnaraamat.* Tallinn: ETA KKI.

Erelt, Mati (editor) 2003. *Estonian Language.* Linguistica Uralica Supplementary Series vol 1. Estonian Academy Publishers, Tallinn.

Evert, Stefan 2005. *The statistics of word cooccurrences : word pairs and collocations*.

URL: **http://elib.uni-stuttgart.de/opus/volltexte/2005/2371/**

Hasselblatt, C., 1990. *Das Estnische Partikelverb als Lehnübersetzung aus dem Deutschen*, Wiesbaden.

Kaalep, Heiki-Jaan, Kadri Muischnek. 2003. Inconsistent Selectional Criteria in Semi-automatic Multi-word Unit Extraction. *COMPLEX 2003, 7th Conference on Computational Lexicography and Corpus Research*, Ed. By F. Kiefer, J. Pajzs, Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest: 27-36

Kaalep, Heiki-Jaan, Kadri Muischnek. The corpora of Estonian at the University of Tartu: the current situation. *Proceedings of the Second Baltic Conference on Human Language Technologies*. Institute of Cybernetics, Tallinn University of Technology. Institute of the Estonian Language. Editors: Margit Langemets, Priit Penjam. Tallinn 2005: 267-272

Katz, Jerrold J. 1973. Compositionality, Idiomaticity and Lexical Substitution. – *A Festschrift for Morris Halle*, ed. By Stephen R. Anderson and Paul Kiparsky: 357-376.

Oflazer, Kemal, Özlem Cetinoglu, Bilge Say 2004. Integrating Morphology with Multi-word Expression Processing in Turkish. *Second ACL Workshop on Multiword Expressions: Integrated Processing*: 64-71.

Nunberg, Geoffrey, Ivan A. Sag, Thomas Wasow 1994. Idioms – *Langue* 70 (3): 491-538

Riehemann, Susanne Z. 2001. *A Constructional Approach to Idioms and Word Formation. PhD dissertation*. URL http://doors.stanford.edu/~sr/sr-diss.pdf

Saareste, Andrus. 1979. *Eesti keele mõistelise sõnaraamatu indeks*. Finsk-ugriska institutionen, Uppsala.

Õim, Asta. 1993. *Fraseoloogiasõnaraamat*. ETA KKI, Tallinn, Estonia.

Õim, Asta. 1991. *Sünonüümisõnastik*. Tallinn, Estonia.

Appendix 1. Abbreviations used in glosses

ABL – ablative case
ADE – adessive case
ALL – allative case
GEN – genitive case
EL – elative case
ILL – illative case
INF – infinitive
PART – partitive case
PL – plural
TR – translative case