Multi-Word Verbs of Estonian: a Database and a Corpus

Heiki-Jaan Kaalep, Kadri Muischnek
University of Tartu
Liivi 2, Tartu, Estonia
E-mail: Heiki-Jaan.Kaalep@ut.ee, Kadri.Muischnek@ut.ee

Abstract

The paper describes two interrelated language resources: a database of 13,000 Estonian multi-word verbs (MWV) and a 300,000 word corpus with annotated MWVs. Both resources have been manually post-edited, and are meant to be used by a wide audience, from corpus linguists to language engineers. The paper gives a short overview of the types of MWVs in Estonian, followed by a description of some grammatical features – word order and inflection – of Estonian and their manifestation in the MWVs. The database is a table that has 13,000 rows and 11 columns and contains information about the source (dictionary or corpus) of the MWV, its linguistic category, frequency in the text corpus, and morphological description. The text corpus contains the morphological analysis of the source text and the annotated MWVs. The layout of the corpus is essentially a table, a row standing for a running word and the columns filled by annotations. The corpus contains 8,200 instances of tagged MWVs and 34,100 simplex main verbs, meaning that roughly every fifth predicate is represented by a MWV. The number of different types of the MWVs in the corpus is 3,500.

1. Introduction

In order to provide an automatic treatment of a language phenomenon, one must first gain a less formal, linguistic understanding of it. Usually it involves making a list of the items one is interested in (morphemes, words, grammar rules etc.) and investigating their behaviour in a real life speech or text corpus. When we are interested in multi-word units (MWU), we face problems that are somewhat similar to those faced by the lexicographers, and researchers interested in morphological analysis and disambiguation: which items should be in the lexicon, how does their form vary, and how do they behave in texts?

Being interested in multi-word verbs (MWV) of Estonian, we have created two interrelated, harmonised resources that complement each other: a database of MWVs and a corpus where the MWVs are tagged. Both of the resources are meant to be used by a wide audience, from corpus linguists to language engineers.

2. Multi-word verbs of Estonian

Estonian belongs to the Finnic group of the Finno-Ugric language family. Typologically it is an agglutinating language. The word order in Estonian reveals remarkable heterogeinity, the written language having tendency towards verb-second pattern. One can find a detailed description of the grammatical system of Estonian in (Erelt, 2003).

In this section we will give a short overview of the types of the Estonian MWVs followed by a brief description of some grammatical features of Estonian posing problems for the automatic treatment of the MWV-s.

In our database we distinguish between the following types of Estonian MWVs:

1. Particle verb (marked *yv* in the database) consisting of an uninflecting particle and a verb (e.g. English *back up*)
2. Expression consisting of a noun (phrase) and a verb (marked as *nv* in the database); could be divided further into idiomatic expressions (e.g. English *kick the bucket*) and collocations (e.g. English *answer the question*).
3. Support verb construction (marked as *sv* in the database) - combinations of a verb and its object (or, occasionally, 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 *take a walk*).
4. Catenative verb construction (marked as *av* in the database) consisting of a verb and an infinitive (e.g. *make do*).

Word order of the MWVs

The heterogenous word order of Estonian means that the components of a MWV can occur in various permutations in a clause and they can be separated from each other by several intervening words as it is the case with the particle verb *üle minema* ‘go over’ in example (1).

(1) *Peavalu **läks** alles järgmisel päeval **üle**.*
Headache go-PST only next-ADE day-ADE over
‘The headache stopped only the next day’

In the examples (2-5) an idiomatic MWV *sõjakirvest välja kaevama* ‘dig out the hatchet, i.e. start the quarrel’ consisting of three components occurs with four different word order variants. In real-life sentences intervening words can occur between all the components of this MWV.

(2) *Jaan **kaevas sõjakirve välja**.*
Jaan-NOM dig-PST hatchet-GEN out
’Jaan started the quarrel’
(3) ***Sõjakirve kaevas välja** Jaan.*
hatchet-GEN dig-PST out Jaan
(4) *Jaan **kaevas välja sõjakirve**.*
Jaan dig-PST out hatchet-GEN
(5) *Kui Jaan **sõjakirve välja kaevas**...*
When Jaan hatchet-GEN out dig-PST

Inflectional variation of the MWVs

Estonian being an agglutinative language means that the verbal component of a MWV inflects freely in texts. In the database it is recorded in its base form and there are

principally two possible ways of matching the database with the texts: either the morphological tagging of the text, or generating all possible forms of the verb in the database.
The non-verbal component of the particle verbs and catenative verbs does not inflect.
However, if a MWV consists of a verb and a NP, the latter may inflect, albeit with various degrees of freedom, which in turn depend on syntactic and semantic features. The rigidness of NPs of MWVs is an important characteristic, and should be recorded accordingly. The MWVs can be divided into subclasses depending on the inflectional behaviour of the nominal component. Among these MWV-s support verb constructions are distinguished as a special subclass. The remaining MWV-s fall into the subclasses of opaque idioms, transparent idioms and collocations. The nominal components of all opaque idiomatic expressions and part of the transparent idiomatic expressions are always frozen in the same case and number and can therefore be treated much like particle verbs in the database.
The flexibility of the nominal components of part of the transparent idioms, most of the support verb constructions and most of the collocations depends on the type of the syntactic relationship with the verb. If the nominal component is formally in the object position of the verb, it can undergo the so-called object case alternations.
Here a few words should be said about the case alternation of the object NP in Estonian in general. Three case forms are possible for the object NP – partitive (both in singular and plural), nominative (singular and plural) and genitive (only singular).
Partitive is the unmarked case form of the object – the 'partial object', as it is often called. The nominative and genitive forms are grouped together under the label 'total object'.
Total object can be found only in an affirmative clause; it cannot be used in a negative clause. The case alternation of the object is used to express the distinction between telic-atelic aspect of the clause. If the verb denotes telic activity (an activity that can have a result), and the activity described in the clause is perfective, then the total object is used:

(6) *Mees ehitas suvilat*
Man built summer-house-PART
'The man was building a summer-house.' (imperfective activity)
(7) *Mees ehitas suvila*
Man built summer-house-GEN
'The man built a summer-house.' (perfective activity)

The nominal components of the transparent idioms are divided 75-25 between the forms of partial object and total object. E.g. in the example (8), the transparent idiom with the nominal component in the form of the total object was used to describe a perfective activity.

(8) *Esinemisele **pani punkti** ilutulestik*.
Show-ALL put period-GEN fireworks
'The fireworks put an end to the show.'

Some of the transparent idioms behave like regular verb-object combinations in this respect, while others show irregular variation, and there are those whose nominal components are frozen in the partitive case. Thus the transparent idioms do not form a homogenous group with respect to the case alternation of the nominal component.
As a practical solution, the information about the variability of the nominal component is recorded separately for each MWV together with the information about the relevant morphological categories (cf sect 3.1).
In support verb constructions, the case alternation of the object is regularly used to express the aspect of the clause:

(9) *Žürii alles **teeb otsust***.
Jury still makes decision-PART
'The jury is still making the decision.' (imperfective)
(10) *Žürii **tegi** lõpuks **otsuse***.
Jury made at-last decision-GEN
'The jury made the decision at last.' (perfective)

Different support verb constructions differ from each other (just like ordinary verbs do) in whether they express an atelic or telic activity. Some support verb constructions are generally used to emphasize the process of the activity (atelic activity), not its result. Such expressions don't normally show case alternation in texts.
In addition to the object case alternations, the nominal components of these three groups can undergo number alternations. Especially the support verb constructions make extensive use of number alternation of the nominal component, whereas the plural form of the noun denoting an action usually refers to several events.

3. Database of MWVs

This database contains multi-word expressions, consisting of a verb and a particle or a verb and its complements. The expressions consisting of a verb and its subject are not included. The multi-word units consisting of a verb and an infinite form of a verb are included irregularly.
The present version of the database contains ca 13,000 expressions.
The database has been compiled on the basis of:
1. Dictionaries and wordlists, aimed at human users, namely:
1.1. Phraseology Dictionary (Õim, 1991),
1.2. The Explanatory Dictionary of Estonian (EKSS 1988-2000),
1.3. Filosoft thesaurus (http://www.filosoft.ee/ thes_et/),
1.4. A list of particle verbs (Hasselblatt, 1990),
1.5. Index of the Thesaurus of Estonian (Saareste, 1979),
1.6. Dictionary of Synonyms (Õim, 1993).
2. The MWVs, extracted automatically from corpora totalling 20 million tokens and post-edited manually. This collocation extraction experiment is described in (Kaalep, Muischnek, 2003).
3. The MWVs found during manual post-editing of the corpus of MWVs (see section 4)

3.1 Database Layout

The database is a table, with every row having 11 fields. The fields are delimited with colons. If a field is empty it means that this information is missing at the moment.

The fields contain the following information:

Field 1
The expression itself. The verbal component of the expression is recorded in the supine form, the traditional form of presenting the Estonian verbs in the dictionaries. As for the expressions consisting of a verb and a noun or a noun phrase, the noun can be 'frozen' in a certain case form or allow certain case alternations. If the nominal component is 'frozen', then it is recorded in the database in this certain case form. If the nominal component can undergo certain case alternations, it is recorded in the database in the partitive case form, but the information about the case alternation is given in the morphological analysis (see field 11).

Field 2
The subtype of the expression. The possible subtypes are:
yv – particle verb
nv – expression consisting of a noun (phrase) and a verb; could be divided further into idiomatic expressions and collocations
tv – support verb construction
av – catenative verb construction

Fields 3-9
Indication that the expression was recorded in a certain dictionary/wordlist and/or was retrieved with collocation extraction methods:
field 3: Phraseology dictionary (Õim, 1991)
field 4: The Explanatory Dictionary of Estonian (EKSS 1988-2000)
field 5: Filosoft thesaurus (http://www.filosoft.ee/thes_et/)
field 6: A list of particle verbs (Hasselblatt, 1990)
field 7: Index of the Thesaurus of Estonian (Saareste, 1979)
field 8: Dictionary of Synonyms (Õim, 1993)
field 9: Automatically extracted collocations

Field 10
If the expression was found and tagged in the corpus of MWVs (see section 4), the number in this field shows the number of its occurrences in the corpus; otherwise, the frequency is zero.

Field 11
Morphological analysis of the expression. This information is needed by programs that tag MWVs in texts: the components of a MWV may be separated by several words, and the form of its components may vary in various ways, depending on the morphosyntactic type of the component and the rigidity of the MWV itself.
The field is delimited by the <morf> and </morf> tags.
The morphological analysis is similar to the one used in the corpus of MWVs.

4. Corpus

A part of a morphologically tagged corpus from http://www.cl.ut.ee/korpused/morfkorpus has been automatically tagged and manually post-edited also for the MWVs. Table 1 shows the composition of the corpus and the number of MWV instances, compared with the number of sentences and simplex main verb instances (auxiliary and modal verbs are excluded from counts). It is worth noting that roughly 20% of all the predicates used in the texts are MWVs.

	tokens	sentences	MWVs	simplex main verbs
fiction	104,000	9,000	3,800	17,000
press	111,000	9,500	2,500	14,500
popular science	98,000	7,300	1,900	12,600
total	313,000	25,800	8,200	34,100

Table 1. Corpus with MWVs tagged.

4.1 Corpus Layout

Here is an example of a sentence '*Nad jätavad ülikooli pooleli*' ('They leave university in-half', i.e. 'They quit the university') containing a MWV '*pooleli jätma*' ('leave in-half', i.e. 'quit'), as it is represented in the corpus:

```
Nad    tema+d //_P_ pl nom //
jätavad    jät+vad //_V_ main indic pres ps3 pl ps af //
                                        #->pooleli jätma#
ülikooli    üli_kool+0 //_S_ com sg gen //
pooleli    pooleli+0 //_D_ //
```

Figure 1: Corpus layout

The text is in 2 columns, delimited by the tabulation character:
1. Wordform and its morphological analysis; this column is actually just a copy from the Morphologically tagged corpus (http://www.cl.ut.ee/korpused/morfkorpus).
2. MWV, surrounded by # and being in a canonical form, i.e. the form used in dictionaries. MWV is situated on the same row with the verbal component. Immediately after the first #, there is an arrow (<- or ->), indicating the direction where the other parts of the MVW are to be found (in our example, the adverb '*pooleli*').
In rare cases, two or more MWVs are tagged on the same row. This happens when the same verb is used in several MWVs at the same time, e.g. *pass out and away.*

4.2 Tagging

Before tagging the MWVs, the corpus had been morphologically analyzed and manually disambiguated (Kaalep, Muischnek 2005). Thus it was possible to automatically tag the candidate MWVs in the texts, according to what could be found in the database of MWVs. It was then the task of a human annotator to select the right expressions, and occasionally to tag new ones, 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.

5. Database vs. Corpus

Table 2 serves to compare the lexicon of MWVs based on the corpus with the entries of the DB.

MWV types in the DB	13,000
MWV types in the corpus	3,500
hapax legomena of MWVs in the corpus	2,100

Table 2. MWV types in the DB and corpus.

The small proportion of MWVs of the DB that can be found in real texts (compare rows 1 and 2) 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 amount of MWS occurring only once in the entire corpus (*hapax legomena*) deserves some explanation.

From the literature, one may find a number of multiword unit (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.

6. Conclusion

This paper described two interrelated language resources: a database of Estonian multiword verbs and a corpus where these expressions are tagged.

The umbrella term “multiword verbs” covers particle verbs, support verb constructions and expressions consisting of a verb and a noun phrase. The latter category encompasses idiomatic expressions as well as collocations.

The database of MWV-s, based on the data of dictionaries as well as collocations extracted from text corpora, contains various types of linguistic information for ca 13,000 expressions.

A corpus of 300,000 words has been tagged for these MWV-s, indicating that roughly one in five predicates is represented by a MWV.

A closer look at the database and corpus indicates that the criteria for selecting MWUs to be included in a database or tagged in a corpus, might actually be in need of reconsideration, taking into account the experience from the field of lexicography.

7. Acknowledgements

The work on tagging MWVs has been supported in 2004-2007 by the national programs “Estonian Language and National Culture” and “Language technology for Estonian”, and by the Estonian Science Foundation.

8. References

EKSS (1988-2000) *Eesti kirjakeele seletussõnaraamat* (A-Žüriivaba). ETA KKI, Tallinn

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

Evert, S. (2005) *The statistics of word cooccurrences : word pairs and collocations*. URL: http://elib.uni-stuttgart.de/opus/volltexte/2005/2371/

Filosoft - Tesaurus. http://www.filosoft.ee thes_et/

Hasselblatt, C. (1990) Das Estnische Partikelverb als lehnübersetzung aus dem Deutschen. Wiesbaden

Kaalep, H-J. Muischnek, K. (2003) Inconsistent Selectional Criteria in Semi-automatic Multi-word Unit Extraction. In *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, pp. 27--36

Kaalep, H-J, Muischnek, K. (2005) 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: 267-272

Saareste, A. (1979) *Eesti keele mõistelise sõnaraamatu indeks*. Finsk-ugriska institutionen, Uppsala

Tael, K. (1988) *Sõnajärjemallid eesti keeles (võrrelduna soome keelega)*. Tallinn: Eesti NSV Teaduste Akadeemia Keele ja Kirjanduse Instituut. Preprint KKI-56

Õim, A. (1993) Fraseoloogiasõnaraamat. ETA KKI, Tallinn

Õim, A. (1991) Sünonüümisõnastik. Tallinn

Abbreviations

ADE – adessive case
ALL – allative case
GEN – genitive case
PART – partitive case
PST – past tense