Using the Text Corpus to Create a Comprehensive List of Phrasal Verbs[1]

Heiki-Jaan Kaalep, Kadri Muischnek

University of Tartu
Tiigi 78-203, 50090 Tartu, Estonia
{hkaalep, kmuis}@psych.ut.ee

Abstract

The paper describes extraction of Estonian multi-word verbs from text corpora, using a language- and task-specific software tool SENVA, which is based on a statistical language-independent software tool SENTA (Dias et al, 2000). The outcome is a comprehensive list of 16,000 phrasal verbs. We describe the extraction tool, manual post-editing principles, and evaluate the outcome in terms of precision and recall, comparing the results with man-made electronic dictionaries, and with the results of a manual extraction experiment of a sub-set of the MWV-s.

.

1. Introduction

In order to be able to analyze and synthesize real sentences of a language, one has to be aware of the common expressions, which may be complicated idioms as well as simple frequent phrases. A special case of such common expressions are verb phrases, i.e. multi-word verbs (MWV-s): phrasal verbs like *ära maksma* (*pay off*) and idiomatic expressions like *härjal sarvist haarama (take the bull by the horns)*. A repository of MWV-s is indispensable for all the levels of linguistic analysis, from POS tagging and syntactic parsing to semantic disambiguation. One can create such a repository from existing dictionaries or other linguistic resources. However, it is well known that multi-word units have been rather neglected by traditional linguistics: the lexicographers have been concentrating on single words while the grammarians have been interested in general wide-coverage grammatical rules. Fortunately, various computational tools have been developed in order to identify and extract multiword units from electronic text corpora on statistical grounds. Using such a tool should enable us to improve the quality of a repository of MWV-s by answering to the following questions:

1. Do texts contain MWV-s that are not in our repository?

2. Are the MWV-s that one finds in our repository used in real-life texts of today?

The extent of the improvement, and the effort needed for it, are, however, difficult to judge beforehand. The only way to find out is to make a large-scale experiment.

An earlier experiment with a software tool called SENVA (Software for the Extraction of N-ary Verbal Associations) (Dias et al, 2001) to extract MWV-s from a 500,000-word corpus of Estonian fiction showed us that quite a number of verbal locutions used in corpus texts were absent from man-made dictionaries. So, in order to build a comprehensive list of MWV-s, we decided to use the same software on a much larger corpus.

2. Database

Our goal was to come up with a comprehensive list of Estonian MWV-s, used in contemporary texts. For building it, we decided to use all the available resources: human-oriented man-made dictionaries, as well as text corpora.

As a first step, we started from the existing dictionaries that were aimed at a human reader, and compiled an electronic database of MWV-s, containing 10,800 entries. We used the following resources: 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 3,000 of the entries were combinations of a verb and an adverbial particle, e.g. *üles võtma* (*take up*); 7,000 were other multi-word verb constructions, notably: verb + noun phrase, e.g. *vande alla panema* (*put under oath*) and verb + verb, e.g. *värisema panema* (*make shiver*).

3. Corpus

If the corpus is big enough and contains a sufficient variety of text types, one should be able to extract an exhaustive list of linguistic phenomena we are looking for. For this task we used 3 different subcorpora of Estonian .

1. Estonian fiction from 1992-1998 (500,000 words). The corpus consists of 2000-word extracts from various authors.

2. Estonian newspapers form the years 1995-2001 (9.8 million words). The corpus contains full numbers of the newspapers, scattered through the period, mostly from the biggest Estonian daily "Postimees" and the biggest weekly "Eesti Ekspress", plus a few numbers from other dailies and weeklies. All these newspapers are nation-wide quality newspapers; the corpus contains no tabloids.

3. Full transcripts of the sessions of the Estonian parliament Riigikogu from 1995-2001 (12.6 million words). This corpus clearly represents written language, not the spoken variety. The parliamentary sessions are not transcribed word for word. Even the first written version of a transcribed session follows the conventions of written language, and the transcripts our corpus contains have been post-edited in addition to that. As for the vocabulary,

[1] We use the term *phrasal verb* here to denote what is *multi-word lexical verb* in English grammars; we use the latter term in the rest of the paper for clarity.

these debates contain a lot of legislative slang, but surprisingly also a lot of idioms and sayings.

The balance between these corpora is far from perfect. We would have liked to have much more fiction texts to analyze, but these are not easily available in the electronic form. It would have been possible to obtain fiction texts from earlier periods, but we wanted to stick to the contemporary ones as the Estonian language has undergone quite a big change during the beginning of the 1990s.

4. SENVA, a tool for finding MWV-s

The task of finding Estonian MWV-s is a difficult one. Estonian is a Finno-Ugric language, the closest relative of it being Finnish. It is a flective language with free word order. Its syntax has, however, been strongly influenced by German, and the usage of phrasal verbs in Estonian is often viewed as being similar to German, characterized by long distance dependencies between words.

Estonian verbs and nouns have tens of different inflected forms. Adverbs, adpositions (pre- and postpositions) and inflectional forms of nouns are often homonymous with each other. Depending on the type of the clause (e.g. main or relative), the order of the components of a multi-word verbal unit may vary. The words of a MWV may be intervened by other words of the sentence.

All this results in a multitude of possible patterns for a single MWV, and consequently in a huge number of collocations that should be evaluated in order to find the linguistically motivated ones.

Thus we face the following tasks:

1. Diminish the set of different word combinations that have to be evaluated.

2. Find the words that occur together more often than they would by chance.

3. From these, find collocations that form MWV-s.

For solving these tasks we have to combine linguistic and statistical methods with manual editing.

Below we will describe the phases of the tool: corpus preparation, collocation extraction and statistical processing.

4.1. Corpus preparation

We take advantage of the fact that in the case of MWV-s, the verb itself may inflect freely, while the other words tend to be frozen forms. So we can diminish the set of different collocations by converting the verbs to their base forms. To do that, we first perform a full morphological analysis and disambiguation of the text. Then we keep the base form for the verbs, mark the verbs for subsequent collocation selection, and retain the original word form for all the other words.

4.2. Collocation selection

Defining a good the set of collocations for subsequent statistical analysis and manual evaluation is of extreme importance. This phase has a profound influence on the precision and recall rates of the tool.

We select all the possible collocations from the linguistically processed corpus, adhering to the following principles.

1. We limit our collocations to 2- and 3-grams, as these are by far the most frequent types of Estonian MWV-s.

2. A MWV cannot cross the border of a clause.

3. Its components should not be further apart than a fixed number of intervening words. We set the number to 0, 1, 2 and 3 in four separate runs of SENVA on the same corpus.

4. We select only collocations with a verb (which we had marked in the corpus preparation phase).

5. We reject collocations that contain certain words that cannot be part of MWV-s:

- proper names
- pronouns (with a few exceptions)
- conjunctions
- auxiliary verbs *olema* (*to be*) and *ära* (*don't*)
- 100 adverbs (e.g. *palju* (*much*), *taas* (*again*))
- 3,000 word forms of nouns that are either too common (e.g. *öösel* (*at night*), *faktid* (*facts*)) or too specific (e.g. *advokaat* (*lawyer*), *arutelu* (*discussion*)) to be part of a MWV.

We created these lists of adverbs and nouns in the following way. First, we extracted all the collocations from a corpus. Then we created a frequency list of single words from this list of collocations, and picked for manual inspection words that were never found in our database of MWV-s. We checked the top of this list and marked the words that we considered impossible to be found in a MWV.

6. Sort the words in every collocation so that the verb will be the last component. Thus the collocations will have the form, used in the dictionaries.

4.3. Statistical analysis

Statistical analysis is performed on the set of collocations, extracted in the previous phase.

The statistical tool we use is SENTA (Software for Extracting N-ary Textual Associations) developed by (Dias et al., 2000) and tailored by us for the specific case of extracting Estonian MWV-s. Below we briefly describe SENTA's underlying principles.

4.3.1. The Mutual Expectation measure (ME)

By definition, multiword lexical units are groups of words that occur together more often than expected by chance. From this assumption, (Dias et al., 2000) have defined a mathematical model to describe the degree of cohesiveness between the words of an *n*-gram.

The normalized expectation (NE) existing between *n* words is defined as the average expectation of one word to occur in a given position, knowing the occurrence of the other *n*-1 words also constrained by their positions. For example, the average expectation of the *3*-gram *take into custody* must take into account the expectation of occurring *custody* after *take into*, but also the expectation of *into* linking together *take* and *custody*, and finally the expectation of *take* to occur before *into custody*. The basic idea of NE is to evaluate the cost of the possible loss of one word in an *n*-gram. The less an *n*-gram accepts the loss of one of its components, the higher its normalized expectation will be. NE is thus defined as the probability of an *n*-gram, divided by the arithmetic mean of the probabilities of *n-1*-grams it contains:

$$NE = \frac{prob(n-gram)}{\frac{1}{n}\sum prob(n-1-grams)}$$

So, the more *n-1*-grams occur somewhere else besides inside the *n*-gram, the bigger the arithmetic mean will be, and consequently, the smaller NE for this particular n-gram will be.

NE can be viewed as a generalization of the Dice coefficient (Smadja 1993), which is equivalent to NE for bigrams:

$$Dice(x,y) = \frac{prob(x,y)}{\frac{1}{2}(prob(x)+prob(y))}$$

From the assumption that one effective criterion for multiword unit identification is simple frequency (Daille 1995), it is posed that between two *n*-grams with the same normalized expectation, the most frequent *n*-gram is more likely to be a multiword unit:

$$ME = prob(n-gram) \times NE(n-gram)$$

When calculating ME for a text corpus containing N running words, the formula, using absolute frequencies, will be:

$$ME = \frac{freq(n-gram)}{N} \times \frac{freq(n-gram)}{\frac{1}{n}\sum freq(n-1-grams)}$$

4.3.2. The GenLocalMaxs Algorithm

Once SENTA has calculated the ME for an *n*-gram, as well as for its *n-1*-grams and shorter -grams contained in it, it uses the GenLocalMaxs algorithm to decide which one among them to choose. The algorithm assumes that an *n*-gram is a multiword unit if the degree of cohesiveness between its *n* words is higher than or equal to the degree of cohesiveness of any sub-group of (*n-1*) words contained in the *n*-gram, and if it is strictly higher than the degree of cohesiveness of any super-group of (*n+1*) words, containing all the words of the *n*-gram. In other words, an *n*-gram, let's say *W*, is a multiword unit if its ME value, *ME(W)*, is a local maximum. If set of the ME values of all the (*n-1*)-grams contained in the *n*-gram *W*, is denoted by Ω_{n-1} and the set of the ME values of all the (*n+1*)-grams containing the *n*-gram *W*, by Ω_{n+1}, then the GenLocalMaxs algorithm is defined as follows in Figure 1.

$\forall x \in \Omega_{n-1}$, $\forall y \in \Omega_{n+1}$
if n=2 then
  if *ME(W)* > *ME(y)*
    then *W* is a multiword unit
else
  if *ME(x)* ≤ *ME(W)* and *ME(W)* > *ME(y)*
    then W is a multiword unit

Figure 1**:** The GenLocalMaxs

For the purpose of our task we found that GenLocalMaxs is sometimes too rigid. When we increased the maximum number of words allowed between the components of a MWV, we got some new good candidates, but also lost some. So we decided to run SENVA on the same corpus for 4 times, each time setting a different limit (0-3) to the maximum number of intervening words. Then we merged the outputs for manual inspection. This increased the likelihood that good MWV-s are extracted, but it also increased the volume of the output.

5. Manual post-editing

We considered everything that passed trough the GenLocalMaxs filter as valid output, even collocations with a frequency of 2 and ME so small that it was shown as zero. Interestingly, such collocations sometimes appeared to be true MWV-s. The decision not to set a frequency or ME-based threshold meant that we had to spend more effort on manual evaluation of the selected collocations.

To select MWV-s from the output, we first grouped it by the verbs. Then one of us browsed through the collocations and marked the ones (s)he considered true MWV-s. This selection was in turn checked by another person.

In the decision process we were guided by the principles, observable in the database of MWV-s, based on human-oriented dictionaries: productively formed MWV-s should not, as a rule, be included.

We discarded the candidate for a MWV when:

1. The verb and its collocate do not form a grammatical MWV, e.g. *aastal hukkuma* (*perish in the year of*)

2. The collocate may be used in conjunction with any verb of the same type, e.g. *asju korraldama, organiseerima jne* (*arrange, organize etc. things*)

3. The verb has a multitude of collocates of the same type, e.g. *aktsiaid, maju jne ostma* (*buy shares, houses etc.*).

4. The collocate is an adverb that acts as a free combination, typically answering the question "how", "when" or "where", e.g. *valjusti* (*loudly*), *äsja* (*recently*), *saalis* (*in the hall*).

We decided to include a phrase if the meaning of the verb was clearly altered by the context. This principle was hard to follow in practice, though. If a function word like an adverbial particle gives the verb a new meaning, then it is straightforward to classify the collocation as a MWV. Content words, in contrast, always add something to the meaning. E.g. *palka saama* (*get a salary*) and *AIDSi saama* (*get AIDS*) have very different meanings, but then, almost everything can be got, so we decided not to include these phrases.

As a rule of thumb, we included phrases were the verb is used non-literally.

Another rule we followed was: if a collocation includes uncommon word(s), it is more likely to get into the list of MWV-s.

6. Results and evaluation

We ran SENVA on three corpora – fiction texts, parliament transcriptions and newspapers.

The following table lists the number of different MWV-s found in the corpora, after manual pruning of the candidate lists that were output by SENVA.

	fiction	parliam.	newsp.
multi-word verbs	3,000	5,800	8,500
of these, in the database	1,900	2,600	4,200
not in the database	1,100	3,200	4,300

Table 1: MWV-s in the corpora.

A considerable part of all the MWV-s found in the corpora was missing from the original man-made dictionaries that had been the sources for our database. These are the MWV-s we were after and finding them in such great numbers justified the whole undertaking.

If we view the MWV extraction results from all the 3 corpora together, we find that they contain 10,900 different multi-word verbs, 4,900 of which had been listed in the database before, and 6,000 of which are new acquisitions from the corpora.

The database we created from human-oriented dictionaries originally contained 10,800 entries, 5,900 of which were not found in any of the corpora. It is highly likely that a lot of them should be eventually discarded from the list of present-day Estonian MWV-s.

The surprisingly large section of the database that could not be found in the corpora may be explained by two reasons. First, much attention of the dictionary-makers has been caught by idioms and sayings that are known to be rare in the (written) language. Second, these dictionaries tend to reflect the language of the fiction and especially the language used in the Estonian fiction up to the eighties of the 20th century, that is before the period we chose our corpus texts from.

6.1. Precision

Using corpora in the size of 10 million words for extracting MWV-s, however, proved to be considerably more laborious than we had expected from the previous experience with a 0.5 million word corpus.

The following table lists the sizes of the corpora, the number of different n-grams SENVA extracted, the number of true multi-word verbs, and the precision.

	fiction	parliam.	newsp.
words (in millions)	0.5	12.6	9.8
extracted n-grams	14,500	272,000	308,000
multi-word verbs	3,000	5,800	8,500
precision	21%	2%	3%

Table 2. Precision in different corpora.

We can see that the growth of corpora by 20 times brought along the increase in the number of extracted n-grams in the same magnitude, while the number of MWV-s increased only 2-3 times, which in turn resulted in a considerable decline in precision.

What we see here is similar to the increase of the size of a corpus, compared with the increase in the size of the wordform lexicon of the same corpus.

We were interested in maximal recall, so we did not make any attempts to diminish the number of n-grams, if it resulted in even a small decline in recall.

6.2. Recall

In order to estimate the recall of SENVA, that is the proportion of all the multi-word verbs in the corpus that SENVA was able to present for linguistic evaluation, we made the following experiment. We selected randomly 500 entries from our database. We checked the corpora manually for these MWV-s, and compared the result with the findings of SENVA. In principle, SENVA can find only phrases that occur at least twice in the corpus. The results are presented in the following table:

	fiction	parliam.	journal.
set of MWV-s	500	500	500
in the corpus	71	130	221
extracted by SENVA	61	107	188
recall	86%	82%	85%

Table 3: Recall in different corpora.

We may conclude from the experiment that 18-14% of the frequently occurring multi-word verbs remain undiscovered by SENVA.

By far the commonest reason for not finding a good candidate lies in the nature of GenLocalMaxs algorithm. If a multi-word verb occurs frequently in a certain context, this wider context will prevail over the shorter. E.g. in the Parliament transcriptions we find that *üles võtma* (*take up*) occurs in the contexts *kutsuma üles võtma* (*call to take up*) and *teemat üles võtma* (*take up a theme*) so often that the 3-grams are selected as candidates for MWV-s, thus neglecting the 2-gram.

The most promising way to remedy this deficiency would be to eliminate bad n-grams on linguistic grounds, e.g. to eliminate n-grams containing both a modal verb and a main verb. This would give the good n-grams a better chance for getting selected by GenLocalMaxs.

7. Conclusion

We set us a goal to build a comprehensive list of Estonian multi-word verbs. We started from dictionaries aimed at human readers and added up the information they contained, resulting in a database of MWV-s. However, this database contained a lot of rarely used idioms and lacked many verbal locutions widely used in texts. So, in order to make a comprehensive list of Estonian MWV-s, we decided to extract verbal locutions from a large text corpus using a language- and task-specific software tool SENVA, which is based on a statistical language-independent software tool SENTA (Dias et al, 2000). This work resulted in a freely available database of Estonian MWV-s, containing 16,000 entries, 6,000 of which were new MWV-s, extracted from the corpora.

8. Acknowledgments

The work has been sponsored by an Estonian Science Foundation grant 4352.

9. References

Daille B., 1995 Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In: J. Klavans and P. Resnik (eds.). *The Balancing Act: Combining Symbolic and Statistical Approaches to*

Language, (pp. 49-66). Cambridge, MA; London, England: MIT Press.

Dias, G., Guilloré, S., Bassano, J.C., Lopes, J.G.P,, 2000. Extraction Automatique d'unités Lexicales Complexes: Un Enjeu Fondamental pour la Recherche Documentaire. In: Christian Jacquemin (ed.), *Journal Traitement Automatique des Langues,* 41/2:447-473.

Dias, G., Kaalep, H-J., Muischnek, K., 2001. Automatic Extraction of Verb Phrases from Annotated Corpora: A Linguistic Evaluation for Estonian. In: *ACL 39th Annual Meeting and 10th Conference of the European Chapter. Workshop: Collocation: Computational Extraction, Analysis and Exploitation,*(pp. 47-53). Institut de Recherche en Informatique de Toulouse and Universite des Sciences Sociales. Toulouse, France.

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

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

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

Smadja F., 1993. Retrieving Collocations from Text: XTRACT. *Computational Linguistics*, 19/1:143-177.

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

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