Automatic Extraction of Multiword Units for Estonian: Phrasal Verbs

Gaël Harry Dias <ddg@di.fct.unl.pt>
Heiki-Jaan Kaalep <hkaalep@psych.ut.ee>

1. Introduction

In order to be able to analyse and synthesise real sentences of a language, it is not sufficient if one knows the words and syntax rules of that language. In addition, one has to be aware of the common expressions, which may be complicated idioms as well as simple frequent phrases.

At present, we don't know much about frequent Estonian expressions. There exist a few publications dealing with such phenomena (EKSS, Hasselblatt 1990, Õim 1993, Õim 1998) aimed at a human reader. Based on these studies, a database of multiword units has been compiled which can be accessed by the following URL:
ww.cl.ut.ee/ee/ressursid/pysiyhendid.html. However, their usage and frequency in real-life texts is still unexplored.

Fortunately, language-independent computational tools have been developed in order to identify and extract multiword units from electronic text corpora (Dias et al. 2000). Their ability to deal with all kinds of languages, and in particular Estonian, is a great motivation to find expressions in real-life texts, and to identify the expressions missing from the database that could enrich it. The procedure is simple: run a statistical program, find expressions among multiword unit candidates, compare the results with the existing database, and add new information.

However, drawbacks are likely to occur: the program may find expressions that make little sense for a linguist, and may fail to find those that a linguist would identify from the text by hand. In order to get most out of a statistical tool, we must take into account the linguistic properties of the text and the expressions we are interested in, as well as the requirements of the statistical tool.

Below we will present a case study to demonstrate a successful way of combining linguistic and statistical processing: extracting Estonian phrasal verbs from a text corpus. We will evaluate the results by comparing them to a database of phrasal verbs, built manually from existing dictionaries beforehand, as well as the database itself.

2. Statistical tool

For the specific case of extracting Estonian phrasal verbs, we tailored a statistical tool SENVA (Software for Extracting N-ary Verbal Associations) that customizes the Software for Extracting N-ary Textual Associations (SENTA) developed by (Dias et al. 2000). SENVA uses a complicated mathematical formula and an algorithm of local maxima to evaluate the degree of cohesiveness between words in a text. Below we briefly describe them --- a more thorough description can be found in (Dias et al. 2000).

2.1. The Mutual Expectation measure

By definition, multiword lexical units are groups of words that occur together more often than expected by chance. From this assumption, we define a mathematical model to describe the degree of cohesiveness that stands between the words contained in an *n*-gram. We use this model to calculate the Mutual Expectation measure, based on the Normalized Expectation.

2.1.1. *Normalised Expectation*

We define the normalised expectation (NE) existing between *n* words as the average expectation of the occurrence of one word 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 "*vahi alla võtma*" (*take into custody*) [*vahi +1 alla +2 võtma*] must take into account the expectation of occurring "*võtma*" after "*vahi alla*", but also the expectation of "*alla*" linking together "*vahi*" and "*võtma*" and finally the expectation of occurring "*vahi*" before "*alla võtma*". This situation is graphically illustrated in Table 1 where one possible expectation corresponds to one respective row.

Table 1: Example of expectations to take into account in order to evaluate the NE

Expectation of the word	**Knowing the gapped 3-gram**
vahi	[_____ *+1 alla +2 võtma*]
alla	[*vahi +1* _____ *+2 võtma*]
võtma	[*vahi +1 alla +2* _____]

The basic idea of the normalised expectation is to evaluate the cost, in terms of cohesiveness, of the possible loss of one word in an *n*-gram. So, the more cohesive a word group is, that is the less it accepts the loss of one of its components, the higher its normalized expectation will be. We define the normalised expectation 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 the text contains *n-1*-grams that occur somewhere else besides inside the *n*-gram, the bigger the arithmetic mean will be, and consequently, the smaller NE will be.

2.1.2. *Mutual Expectation*

Daille (1995) shows that one effective criterion for multiword unit identification is simple frequency. From this assumption, we pose that between two *n*-grams with the same normalised expectation, that is, with the same value measuring the possible loss of one word in an *n*-gram, the most frequent *n*-gram is more likely to be a multiword unit:

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

So, the Mutual Expectation (ME) between *n* words is based on the normalised expectation and the relative frequency. Once we have calculated the ME for an *n*-gram, as well as for its *n-1*-grams, *n-2*-grams and shorter -grams contained in it, we face the following question: which one among them to choose?

2.2. The GenLocalMaxs Algorithm

To answer the previous question, we use the GenLocalMaxs algorithm. The GenLocalMaxs elects the multiword units from the set of all the cohesiveness-valued *n*-grams based on two assumptions. First, the more cohesive a group of words is, the higher its ME score will be. Second, multiword lexical units are highly associated localized groups of words. From these two assumptions, we may deduce that an *n*-gram is a multiword unit if the degree of cohesiveness between its *n* words is higher or equal than 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. As a consequence, an *n*-gram, let's say *W*, is a multiword unit if its ME value, *ME(W)*, is a local maximum. Let's define the set of the ME values of all the *(n-1)*-grams contained in the *n*-gram *W*, 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}. 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 $ME(x) \leq ME(W)$ and $ME(W) > MEl(y)$ then W is a multiword unit

Figure 1: The GenLocalMaxs

3. Text preparation

Estonian is a flective language with a free word order. It belongs to the Finno-Ugric family, the closest relative being Finnish. 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 frequent use of long distance dependencies between words.

Due to the nature of Estonian, it is likely that a language independent statistical tool will perform poorly: the program may find expressions that make little sense for a linguist, and may fail to find those that a linguist would identify from the text by hand. The reason for this drawback is that statistical systems cannot differentiate between important and unimportant variability in texts, thus failing to recognize similar patterns. Indeed, they are designed to identify recurrent and probable associations between wordforms and do not take advantage of the specificities of the language. Indeed, it is likely to find inflectional endings that may weaken the results of the extraction. So, it is natural to assume that inflectional endings of the components of an expression in real text may present important variability. Eliminating them, by using a lemmatiser, would give the statistical software better grounds for finding recurring patterns. Consider, for example, expressions like "*saalomonlik otsus*" (*Salomon's decision*) where both components may freely inflect. Giving up the inflectional endings would provide great benefits to the process of extraction.

However, at the same time, it is known that expressions tend to contain frozen forms, including inflectional endings, and eliminating them might lose information, necessary for recognizing the expression. For example, in "*hullu lehma tõbi*" (*mad cow syndrome*), one may never use any other form, like "*hull lehm*" (*mad cow*, singular nominative case) or "*hullude lehmade*" (*mad cows*, plural genitive case) instead of "*hullu lehma*" (*mad cow*, singular genitive case) in the context of "*tõbi*" (*syndrome*). Correspondingly in English, one may not say "*Human Right*" or "*Humans Right*". Instead, one must always take into consideration the

inflection of both the constituents and produce "*Human Rights*" as the only correct expression.

Phrasal verbs like "*ära maksma*" (*to pay off)* and idiomatic verbal expressions like "*end tükkideks naerma*" (*to laugh oneself to pieces)* represent a situation that is different from both of the above-mentioned extremes: the verb part may inflect freely, but the other word(s) are frozen forms, and the order of the constituents of a phrasal verb may vary, according to the type of the sentence where it occurs. Consequently, we tried a pragmatic approach to text preparation: lemmatise only some words (the ones that inflect freely in the expressions), and do not lemmatise others.

4. Experiment

We made our experiment on a 500,000-word sub-corpus of the Corpus of Written Estonian of the 20th Century (-bin/konk_sj_en.cgi). The corpus we use is similar to the classical Brown corpus. It consists of 2,000-word long extracts from original Estonian prose from 1992-1998. We chose a prose corpus for our experiment because written prose (as opposed to newspapers, poetry, or spoken language) has always been the main source of inspiration for "traditional" linguists and dictionary makers. We expected that by using the same type of text that the traditional dictionary makers, we could better compare our database (it is based on traditional hand-made dictionaries) with the output of the computational tools we use.

The database of Estonian phrasal verbs contained 10816 entries. It was based on the following sources, originally aimed at a human reader: (EKSS), (Saareste 1979), (Hasselblatt 1990), (Õim 1991), (Õim 1993), (Filosoft). Only two- and three-component phrasal verbs were included. So, in order to extract phrasal verbs, we performed the following tasks.

1. Perform a morphological analysis and disambiguation of the corpus.
2. For verbs, keep the lemma form; for other words, keep the original wordform.
3. Select all the possible collocations.
4. Eliminate collocations, not relevant for this particular task. That is, eliminate collocations not including a verb, as well as collocations containing pronouns (with a few exceptions), punctuation, certain adverbs etc.
5. Calculate Mutual Expectation and GenLocalMaxs; based on these, make the final choice of extracted phrases.

We processed the corpus four times with SENVA, each time setting a different limit (0 to 3) to the number of words that may intervene the words that belong to a phrase. Then we combined the results and compared them against the database we had. We checked manually all the extracted phrases that were not in the database, and decided whether they should be added or not.

5. Results

SENVA extracted 13,100 phrasal verb candidates. 2,500 of these, 19%, are such that they should be found in a database of Estonian phrasal verbs. The rest are collocations a linguist would rather not present in the database. In fact, 1629 of the 2,500 were expressions that the database already contained; and SENVA found extra 865 phrases that should be included. Table 2 presents some phrasal verbs that were in the database and/or found by SENVA, illustrating the lack of commonplace expressions from the database, and the existence of rare expressions at the same time.

Table 2: Some phrasal verbs in the database and/or text corpus

Phrase	**In the DB**	**Found by SENVA**
abiellu astuma (to marry)	+	-
abiellu heitma (to marry)	+	-
abielu rikkuma (to commit adultery)	+	-
abielu sõlmima (to contract a marriage)	+	-
abielu lahutama (to divorce)	-	+
andeks andma (to forgive)	+	+
andeks paluma (to apologize)	+	-
andeks saama (to obtain forgiveness)	-	+
allkirja andma (to give one's signature)	-	+
hulluks minema (to go mad)	+	+
hulluks ajama (to drive mad)	-	+
külla minema (to go on a visit)	+	-
külla tulema (to come on a visit)	+	+
külla kutsuma (to invite)	-	+

The figures 1629 and 865 give us an estimation of the quality of our database. The database contained 10816 phrasal verbs, which is quite a lot. However, we can see that out of the 2,500 phrasal verbs that were extracted from the corpus, only 2/3 were represented in the database also. Without SENVA, we would not have been able to find the missing 1/3.

6. Evaluating SENVA

How sure can we be that SENVA really found all the phrasal verbs that are in the corpus, and that it did not report about phrasal verbs that are really not there? For estimating this, we made an experiment with 500 randomly selected phrasal verbs that we had in our database before. By checking the corpus manually, we found that 131 out of the 500 could be found in the corpus. In principle, SENVA can find only phrases that occur at least twice in the corpus. The number of such phrases was 71 (out of the 500).

We made 4 experiments with SENVA, where we defined differently the number of words that could possibly occur between the words of a phrase: 0, 1, 2 or 3. We also combined the results of all the four experiments.

Table 3: Number of correct phrases found by SENVA

Distance	0	1	2	3	combined
Phrases	45	46	50	52	57

From the table above we can see that the longer the allowed distance between individual words of a phrase, the more possible phrases SENVA finds. It is noteworthy, however, that as the distance grows longer, SENVA stops finding some phrases that it did with a shorter distance. A sudden change happens when we change the distance from 2 to 3: SENVA stops finding phrases that occur very often in the corpus. Among the 19 phrases that it did not find, 12 occurred in the corpus twice, but 5 were rather frequent (see table 4).

Table 4: Distance 3: frequently occurring missed phrases

Co-occurring words	Co-occurrences	Phrases
ette näitama (to demonstrate)	10	9
hakkama saama (to be able to cope with)	95	58
suitsu tegema (to have a smoke)	11	9
ära kasutama (to make use of, to exploit)	21	19
ära maksma (to pay off)	12	9

When we used distance 0, 1, or 2, SENVA failed to find only the phrase "*ära maksma*" (*to pay off*), out of the 5 frequent phrases above (although it failed to find more low-frequency phrases than in the case of distance 3).

Although SENVA may have failed to find phrases that are in our database, it often found phrases that contain the ones in our database. Let us consider the phrase "*ära maksma*" (*to pay off*). SENVA was able to find two phrases that contain it: "*arve ära maksma*" (*to pay the bill*) and "*võlga ära maksma*" (*to pay off the debt*). When evaluating SENVA in the current experiment, we did not give it any credit for such findings, although we added both phrases to our database.

Actually, SENVA has pointed to an error in our database: the phrasal verb "*ära maksma*" (*to pay off*) is always used in conjunction with "*arve*" (*bill*), "*võlg*" (*debt*) and a few other nouns, so that the shorter form should be discarded as a phrasal verb.

In addition to the correct phrases, SENVA also found some other phrases (out of the 500) that occur in the corpus. However, these extra phrases were extracted erroneously. What do we mean by "erroneously"? To understand it, we must remember that it is possible that all the words of a phrase co-occur in the same sentence, without, however, forming this phrase in that particular sentence. Let us consider "*tagasi tegema*" (*to pay back*) in the context "*tagasi jõudes teeme sotid selgeks*" (*we will pay when we get back*). SENVA extracted the phrase "*tagasi tegema*" (*to pay back*), and this was an error. It is true that SENVA might accuse us for being unfair --- after all, you should not expect a statistical tool to understand a sentence --- but we just wanted to know how different is SENVA from a human linguist in this respect. Just as one might expect, the number of mistakenly extracted phrases grew with the distance we allowed between the words, reaching 15 (7 that never occurred in the corpus plus 8 that occurred once), if the distance was 3.

What are the conclusions from the experiment with the 500 phrases? We assume that the 131 phrases we found from the corpus form a random selection from all the phrases that are in the given corpus. SENVA could possibly have found only those that occur at least twice, i.e. 71 of them. Actually, in the best possible case, if we combine the results of the experiments with different distances, we may assume that of all the possible phrasal verbs that are in the corpus in the first place, SENVA will find 57/71=80% of those that occur more than once (and close to 99% of those that occur more that 3 times), and 8/60=12% of those that occur once. This evidences a very high recall rate thus balancing the lower precision results.

If a linguist processes a new corpus with SENVA, and picks out only the linguistically good-looking phrases (in our combined experiment 57+7+8=72), (s)he may expect that the final result actually consists of the following:

57/72=79% phrases that occur in the corpus twice or more
8/72=11% phrases that occur in the corpus once
7/72=10% phrases that do not occur in the corpus at all

7. Conclusion

Extracting phrasal verbs from a corpus, especially when one does not have a list of possible candidates at hand, is a difficult task. We have seen that one can accomplish it by combining automatic linguistic and statistical processing with manual post-editing.

Although the precision rate of the extracted phrasal verbs was low, it was compensated by a high recall rate. This in turn means that SENVA is a useful tool for lexicography: browsing 13,100 candidates for verb phrases is very different from browsing the 500,000-word corpus for the same verb phrases.

An unexpected result concerns the evaluation of the database: our database is far from perfect, but so are the dictionaries it is based upon! The results obtained by using SENVA are immediately usable in linguistic processing of Estonian: without a good list of phrasal verbs, even syntactic analysis would be much harder to perform, not to speak about semantic analysis.

Acknowledgments

The research has been partly financed by the Estonian Science Foundation grant 4352.

References

Daille B. (1995) *Study and Implementation of Combined Techniques for Automatic Extraction of Terminology*. The Balancing Act: Combining Symbolic and Statistical Approaches to Language. Ed. by J. Klavans and P. Resnik. Cambridge, MA; London, England: MIT Press, 49-66

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*, Journal Traitement Automatique des Langues, Vol 41:2, Christian Jacquemin (ed.). Paris, France, 2000, pp 447-473.

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

Filosoft – *Tesaurus*

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

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

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

Õim, A. (1998) *Väljendiraamat.* Eesti Keele Sihtasutus, Tallinn, Estonia.

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