Inconsistent Selectional Criteria in Semi-automatic Multi-word Unit Extraction

Heiki-Jaan Kaalep, Kadri Muischnek

University of Tartu
Estonia

{hkaalep, kmuis}@psych.ut.ee

The paper describes an experiment of finding Estonian multi-word verbs in a text corpus. After describing the used method, the paper goes into more detail in describing: first, the extraction of a set of multi-word collocations, second, the criteria for manual selection of the good candidates from this set, and third, the precision and recall of the proposed method.

Introduction

It is widely recognised that knowledge of multiword expressions is indispensable for analysing and producing natural language. A special case of such multiword expressions is multi-word verbs (MWV-s): phrasal verbs like *ära maksma* (*pay off*), idiomatic expressions like *härjal sarvist haarama (take the bull by the horns),* and semantically light support verbs like *tähele panema* (*pay attention to*). 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. So, one has to resort to extracting the MWV-s from text corpora, making the best use of existing software.

In (Kaalep, Muischnek 2002) we reported about an experiment, involving the creation of a database of Estonian MWV-s, based on both human-oriented dictionaries and various text corpora. The current paper has a closer look at one subtask of the experiment – finding new MWV-s in a corpus. We use a statistical method to find these multiword units from the corpora, extracting the MWV-s as collocations.

For practical reasons, we limit our collocations to bi- and trigrams, because these make up the vast majority of Estonian MWV-s.

1 What is a MWV?

What kind of verb-adverb/noun collocations, that is, arbitrary and recurrent word combinations (Benson 1990) should be included in a lexicon that is used in a broad range of applications? Do we need only the MWV-s, which meaning isn't predictable from the meanings of its components? Or those that make up a translation unit? We decided to use a broad definition of the phenomena we are looking for. We took the lexicographic point of view: a MWV is a word combination that should be presented in a dictionary. So we took into account the frequency of a MWV as well as the opacity of meaning.

Estonian is a highly flective language with a free word order. A verb or noun has normally tens of different inflectional forms. The parts of a multi-word verb may be separated from each other by several intervening words. The order of the parts of a multi-word verb is different in different sentence types. To make matters worse, we must note that a collocation has often been described as a co-occurrence of two elements. This definition doesn't fit our objectives, because a MWV in Estonian may consist of more than two words.

However, Estonian syntax is in many ways similar to German. In fact, the grammatical system of Estonian MWV-s, as well as many of the actual MWV-s, has been borrowed from German.

The following types of MWV-s are distinguished:

1) phrasal verbs (combinations of a verb and an adverb). The meaning of such a combination can be the sum of the meanings of the separate words (*come in*) or the phrasal verb can have its own meaning (*grow up*). The Estonian grammatical tradition regards the phrases of the first type as MWV-s as well. They are also presented in our database, because a) it's the easiest way to treat them and b) one can predict that not all theoretically possible combinations are actually used.

2) Idiomatic or figurative combinations of noun(s) and a verb

3) Support verb constructions, i. e. the combinations of noun(s) and a verb, where the nominal part contributes the major part of the meaning to the construction. We didn't restrict the support verbs to the case of semantically light verbs, nor to the syntactic pattern verb + deverbal noun (cf. Calzolari et al (2002) p. 1935).

Because of the vagueness of the semantic criteria, the post-editing of the support-verb constructions was the most complicated part of the manual work. Evert and Krenn (2001, pp 40-41) also consider the borderline between German figurative and support-verb constructions fuzzy. On the other end of the spectrum, the line between support-verbs and free word combinations is also vague.

2 Mutual Expectation and LocalMax

Multiword lexical units are groups of words that occur together more often than expected by chance. From this assumption, mathematical models have been defined to describe the degree of cohesiveness between the words of an *n*-gram, and to calculate the associacion measures (AM) between words, e.g Mutual information, t-score, log-likelihood, χ^2 etc. We decided to use Mutual Expectation (ME) and LocalMax, as described by Dias and Guilloré (1999). It is suitable for finding 2- and 3-grams simultaneously and requires no pre-set threshold.

Below we give a simplified overview of ME. The full mathematical description can be found in Dias et al. (2000).

ME is based on Normalised Expectation (NE), which is a generalisation of a widely known AM, the Dice coefficient (Smadja (1993)). For 2-grams (x, y), NE and the Dice coefficient are equal:

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

NE existing between *n* words is defined as the average expectation of a word to occur, knowing the occurrence of the other *n*-1 words in the vicinity of this word. The basic idea of NE is to evaluate the cost of the possible loss of one word in an *n*-gram. The more the words are interdependent, i.e. 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 :

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

The more an *n-1*-gram occurs 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.

From the assumption that one effective criterion for multiword unit identification is simple frequency, 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:

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

Once we have calculated the ME for a 3-gram, as well as for the 2-grams it contains, we use the LocalMax algorithm (Silva et al. (1999)) to decide which one among them to choose. For 2- and 3-grams, the algorithm is extremely simple:

if ME(2-gram) > ME(3-gram), choose 2-gram
if ME(2-gram) ≤ ME(3-gram), choose 3-gram

3 Existing resources: DB and corpora

Prior to starting extracting collocations from corpora, a database (DB) was created, based on six human-oriented dictionaries and containing 10,800 MWV-s. 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 Filosoft thesaurus (http://ee.www.ee/Tesa/). 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*).
The database is available from http://www.cl.ut.e
Presentation of the MWV-s in the Estonian dictionaries appeared to be somewhat inconsistent: e.g. a phrase *hulluks minema (to go mad)* was present in several human-oriented dictionaries, but the phrase *hulluks ajama (to drive mad)* was absent from all of them.
Nevertheless, we decided to use the database for evaluating from a lexicographic perspective the suitability of extracted collocations, both the ones that were in the DB already, and by similarity, the ones that were new. We assumed that although the dictionaries are not perfect, they nevertheless embody the practical choices the creators had to make, and thus serve as an example we should follow in our own decisions.
Actually, the 6 dictionaries are very different from each other, in respect of what MWV-s the authors have considered worthy of including. The most characteristic features of the dictionaries are the following.
Dictionary of Phrases (Õim, 1991) contains many figurative, rare expressions, e.g. *lubadustega toitma* (*feed with promises*), *tuuleveskite vastu võitlema* (*fight the windmills*).
Index of the Thesaurus of Estonian (Saareste, 1979) contains many ordinary, frequent, non-figurative expressions, rather similar to free word combinations, e.g. *puhkust veetma* (*spend a vacation*).
The list of particle verbs (Hasselblatt, 1990) contains many ordinary, frequent expressions, because they are interesting in comparison with German, e.g. *ette nihutama* (*push forward*), *alla libisema* (*slip down*).
Dictionary of Synonyms (Õim, 1993) contains many free word combinations that the author has included, because they happen to have a one-word synonym, e.g. *halba nalja tegema* (*make a bad joke*) as a synonym of *pilkama* (*mock*).
These differences in the principles of including expressions in the different dictionaries mean that we cannot regard the database as a model of what should be included and what should be rejected.
We used three different corpora:

1. 0.5 million words of fiction from 1992-1998.
2. 9.8 million words of newspapers form 1995-2001.
3. 12.6 million words of transcripts of the sessions of the Estonian parliament Riigikogu from 1995-2001.

4 Extracting Verb Collocations

The MWV-s were extracted separately from all the three corpora. The text was always processed as follows.
1. Morphological analysis and disambiguation with a bigram HMM tagger were performed, using *estyhmm* as described in Kaalep and Vaino (2001).
2. Verbs were labelled and their lemma form was kept; for other words, the original wordform was kept.
3. All the possible 2- and 3-grams, containing a verb and not crossing the clause border, were extracted. The number of intervening words between the components was not important, just as their exact order, so a maximal possible span (in words) was set and some freedom in the order was allowed. E.g. if the span is set to zero, and a verb is in position x, then the following combinations were allowed: (x-1, x), (x, x+1), (x-2, x-1, x), (x, x+1, x+2). With span set to two, the following combinations were added: (x-3, x), (x-2, x), (x, x+2), (x, x+3), (x-4, x-3, x), (x-3, x-2, x), (x, x+2, x+3), (x, x+3, x+4). One can see that not all the possible 3-grams were allowed, e.g. (x-1, x, x+1) or (x-3, x-1, x). Intervening words were allowed only between the verb and its nearest collocate, and the collocates both had to be on the same side of the verb. Thus the collocates in a 3-gram were viewed like a single word, having a space in it.
4. Unsuitable verb collocations, that is, containing impossible words for a MWV, were deleted. The "bad words" were:
- 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.

These lists of adverbs and nouns were created in the following way. First, all the collocations were extracted from a corpus. Then a frequency list of single words from this list of collocations was created, and words that were never found in the database of MWV-s, were picked for manual inspection. The top of this list was checked and the words that were considered impossible to be found in a MWV, were marked.
This stage is of extreme importance for diminishing the manual effort of browsing the extracted collocations, and more sophisticated methods would actually be needed.
5. The words in every collocation were sorted, so that the verb will be the last component. Thus it was possible to bring together the instances of MWV-s that contain the same words, but in a different order, and present them in the form, used in the dictionaries.
6. ME for every collocation was computed.
7. Using LocalMax, the final selection between the collocations was made, and the result was presented for manual inspection. ME and LocalMax were used only for choosing between 2-grams and 3-grams

that contain them. E.g. from *sõnu tähele panema* (*take notice of the words*) and *tähele panema* (*take notice*), the 2-gram should be chosen.
Neither ME nor plain frequency was used for ordering the collocations. We could not find a threshold, discriminating the set of clearly more probable good collocations from the set of less probable ones.
Each of the three text corpora was processed four times, every time with a different maximal span (0-3) between the words of a collocation. The reason behind this is purely pragmatic. As a rule, the longer the maximal span, the more possible collocations can be extracted from a corpus, and consequently, the more MWV-s also. It turned out, however, that increasing the maximal span resulted in losing some of the MWV-s from the set of extracted collocations. This happened because with a longer span, besides the set of "good" collocations, the set of "bad" collocations may increase as well, and LocalMax will delete some good n-grams, because bad n-1-grams or n+1-grams appear to be more frequent.
Being interested in finding as many MWV-s as possible, we just merged the results from all the four runs with different spans, for manual inspection.

5 Manual selection of good MWE-s: conflicting selectional criteria

We didn't set any frequency constraints to the extracted collocations and this decision meant that we had to spend more effort on manual evaluation of the selected collocations. All the extracted collocations were divided between two human evaluators, who browsed them and decided which collocations to include. The outcome of one evaluator was sporadically checked by the other, the differences in decisions discussed, and the policy harmonized. So the final DB is to a certain degree inconsistent, just as the six human-oriented dictionaries are.
Manual post-editing revealed more problems than we expected. With a huge amount of statistically relevant collocations, we had to make uniform and consistent decisions whether to include them in our database or not. In doing so, we attempted to follow the criteria, set forth by creators of human-oriented dictionaries. These criteria, however, appeared to be governed by the specific aims of the dictionary-makers, sometimes out of line with each other.
As a rule of thumb we followed the principle, that a regular free combination of a verb and a noun/adverb should not be included in the DB. (The most frequent free adverbials had been already included in the list of "bad collocates" during collocation extraction phase, and thus excluded from the collocations.) This is a practical lexicographic criterion.
We decided to include:

1) All the extracted phrasal verbs, mostly consisting of an adverb showing direction/ orientation (e.g. *üles tõusma (get up), üles minema (go up)*. This is in line with Estonian grammatical tradition, thus representing a "historical" criterion.
2) All the figurative constructions. The corpora contained fewer figurative MWV-s than one would expect, given their remarkable presence in dictionaries. Newly coined figurative constructs are considered to be the interesting outcome of our venture, so here we followed the "psychological" criterion.
3) "Cranberry collocations", i.e. collocations that include uncommon word(s) and are thus more likely to get into the list of MWV-s. E.g. we included *õlgu kehitama (shrug shoulders*) because *kehitama (shrug*) is almost never met in any other context.

While choosing among the possible support-verb constructions, we started from the principle, that a construction, where a collocate can be used with many verbs belonging to the same semantic class, should not be included in the DB. Likewise, if a verb has several collocates of the same type, the collocations are formed regularly and thus should not be included in a DB. However, as the support-verbs also tend to be used with semantically related collocates, the borderline between support-verb constructions and free combinations is hard to draw and questionable decisions are bound to be made. In view of the above, one should be cautious when (s)he evaluates the outcome in terms of precision and recall.

6 Results and evaluation

The following table lists the number of good collocations that were chosen as MWV-s during manual post-editing.

Table 1. MWV-s in corpora and human-oriented dictionaries

	3-grams	2-grams	total
Only in corpora	1000	4300	5300
Only in DB	2700	4000	6700
Both in the corpora and DB	300	4300	4600

A considerable part of all the MWV-s found in the corpora was missing from the database. On the other hand, the database contained 6700 entries that were not found in corpora. The dictionaries pay a lot of attention to the idioms and phraseological expressions that are known to be rare in the (especially non-fictional) written language. Second, the dictionaries that our DB was based on, reflect mostly the language of the first half of the 20th century, but the corpora date from the 90ties.

6.1 Precision

According to the generally recognized practice, one should present numeric values of the precision and recall. Table 2 shows the sizes of different corpora, the amount of simplex verbs in them, the amount of extracted collocations, the amount of MWV-s (in thousands) and the precision.

Table 2. Verb types and precision in different corpora.

	Fiction	Parliam.	Newsp.
Words in corpora	500	12 600	9800
Instances of simplex verbs in corpora	112	2566	1823
Different types of simplex verbs	3.4	3.4	6.3
Extracted collocations	14.5	272	238
Different types of MWV-s	3	5.8	8.5
Precision	21%	2%	4%

The precision looks quite low at the first glance, especially in comparison with the results of Baldwin and Villavicencio (2002), who have achieved nearly 90% precision with a 1-million-word corpus; that is a much higher percent than 12.3 % that Krenn and Evert (2001) have achieved extracting German figurative verb and support-verb constructions from a corpus containing 8 million words of newspaper texts. They showed that among the 10400 different verb+particle combinations that had the frequency f ≥ 3, 1280 or 12.3% were the good collocations. Our experiment shows that 8500 or only 4% of the 238

000 collocations (both 2- and 3-grams) extracted from a newspaper corpus consisting of 9.8 million words and having the frequency f ≥ 2, were good collocations or MWV-s worth including in a database. So drastic difference in the precision might be explained by the imposed limit on the number of possible collocates of the verb. Restricting the set of possible collocates is a very effective way of achieving better precision. Baldwin and Villavicencio (2002) allowed only 73 particles as possible collocates, while Krenn and Evert didn't limit the possible collocates. However, if we stick to our goal – to find all possible MWV-s (phrasal verbs, figurative constructions and support verbs) from the corpora, we can't limit the set of possible collocates neither by setting a high threshold of frequency or AM, nor by using a pre-determined set of collocates. (Our final database of MWV-s contained 5800 different collocates.)

The fact that the ratio between the corpus size, or the number of verb instances, and the number of extracted collocations stay the same in different corpora, means that the increase in the corpus size results in a considerable increase in the amount of manual work.

It is noteworthy that the corpora are quite different: between the two corpora more or less equal by size, the parliamentary transcripts contain considerably less different verbs than the newspaper texts. The amount of different verbs in the corpus of parliamentary debates is nearly the same as in the corpus of newspaper texts that is 25 times smaller in size. This should at least partially explain the twofold difference in precision between the results using the Parliamentary debates corpus and the newspaper corpus.

The human factor plays also important role in interpreting the percents of precision. The final selection of the good collocations was done by humans. First from a 0.5 million fiction corpus 14,500 collocations were extracted, among these quite a number of verb-noun pairs that could be viewed as free combinations. As the dictionaries that were taken as a golden standard, also contained some free combinations, the human post-editors tended to choose these as good candidates. From a 10-million word corpus we got 20 times more possible candidates. If the collocations come in great numbers, the productive patterns stand out more clearly and the humans tend not to choose free combinations as good candidates. In addition the parliamentary debates contain a lot of legislative slang and MWV-s used only in that sublanguage. The specialized nature of these could influence the humans not to choose them.

Considering all this, the precision characterizes primarily the research task, then the properties of the used corpora, and finally the preferences of the human post-editors. One should be careful in interpreting precision as a measure of the quality of the chosen collocation extraction method.

6.2 Recall

In order to estimate the recall, that is the proportion of all the MWV-s in the corpus that were presented for linguistic evaluation, we made the following experiment. We selected randomly 500 MWV-s from the DB, checked the corpora manually for these MWV-s and compared the results with the extracted collocations.

The experiment showed us that 18-14% of the MWV-s occuring at least twice remain undiscovered. By far the comonnest reason for that lies in the nature of LocalMax algorithm. If a MWV occurs frequently in a certain context, then 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 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, this would give the good n-grams a better chance for getting selected by LocalMax.

7 Discussion

The use of ME and LocalMax for finding collocations is like applying a poorly performing parser to find words related to verbs; the extracted collocations are like branches of a syntactic tree structure, so we can get by without a syntax analyser.

There are lots of AM-s that different authors have used to measure the strength of the connection between the words: MI, Dice coefficent, $\chi 2$, t-score, log-likelihood. Krenn and Evert (2001) have shown that none of them proved to be significantly better than the others while extracting German support verb constructions and figurative expressions that are close by nature to these MWV-s we were looking for. We can't state that using ME together with LocalMax gives us better results than some other AM, but the advantage lays in the possibility to extract both 2- and 3-grams.

The amount of the collocations that a human post-editor has to consider affects his or her choices. The linguistic preprocessing of the texts and the exact criteria for post-editing the extracted collocations have a major influence on the precision and recall rates.

J. Sinclair has pointed out that while doing corpus linguistics, one should be guided by the corpora, not by the linguistic prejudice or concepts. We have done just the opposite – we have used the corpora to look for a phenomenon prescribed by linguistic criteria: MWV-s that could be presented in a lexicon. Maybe our lexicographically oriented task was ill-formulated? Perhaps we can't get a full list of multi-word verbs, because this is an productively formed set and should be presented using a (restricted) grammar, like the set of possible word forms is presented, using word stems and formation rules?

We made a point of not including productively formed free combination collocations in our database as MWV-s. However, they might be needed in several practical applications like e.g. machine translation, so including them might be a good idea after all. Needless to say, this would have a profound influence on the precision rates of the method.

Conclusion

It is not easy to compile a list of multi-word verbs for a language. The human-oriented dictionaries contain lots of rarely used idioms and lack many MWV-s widely used in texts. So the list based on the dictionaries should be enriched with information extracted from text corpora. We decided to extract multi-word verbs from a big text corpus of Estonian containing more than 20 million words. This work resulted in a database of multi-word verbs that contains 16600 entries, 5300 of which had not been included in the dictionaries, but were found in the corpora. Our method is characterised by low precision and a large amount of manual work, that increases with the growth of the size of the used corpora. Considering the task however, we find them inevitable. We also found that dictionary creators have different motives for including MWV-s in their dictionaries, and when selecting entries for a general-purpose MWV DB, one has to take into consideration several inconsistent criteria. We concluded that linguistic preprocessing of texts and defining the exact criteria for manual selection of the extracted collocations, but also the sublanguage and size of the used corpora have an enormous impact on the precision rates, as well as on the amount of manual work.

References

Baldwin, T. and Villavincencio, A. (2002) *Extracting the Unextractable: A Case Study on Verb-particles*. In: Proceedings of the Sixth Conference on Computational Natural Language Learning (CoNLL 2002), Taipei, Taiwan, pp. 98-104

Benson, M. (1990) *Collocations and general-purpose dictionaries*. International Journal of Lexicography 3(1), pp 23-35.

Calzolari, N., Fillmore, C., Grishman, R., Ide, N., Lenci, A., MacLeod, C., Zampolli, A. (2002) *Towards Best Practice for Multiword Expressions in Computational Lexicons*. In: Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002), pp 1934-1940.

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

Evert, S., Krenn, B. (2001) *Can we do better than frequency? A case study on extracting PP-verb collocations*. In: Proceedings of the ACL/EACL 2001 Workshop on the Computational Extraction, Analysis and Exploitation of Collocations, pp 39-46.

Dias, G. and Guilloré, S. (1999) *Multilingual Aspects of Multiword Lexical Units*. In: Proceedings of the Workshop: Multilingual Language Technologies, Ljubljana, pp. 11-21

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, 41/2, pp 447-473

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

Kaalep, H-J., Muischnek, K. (2002) *Using the Text Corpus to Create a Comprehensive List of Phrasal Verbs*. In: Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)), pp. 101-105.

Kaalep, H-J. and Vaino, T. (2001) *Complete Morphological Analysis in the Linguist's Toolbox*. In: Congressus Nonus Internationalis Fenno-Ugristarum Pars V, Tartu, pp. 9-16.

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

Silva, J.F., Dias, G., Guilloré, S., Lopes, G. P. (1999) *Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units*. In: Progress in Artificial Intelligence, vol. 1695 of Lecture Notes in Artificial Intelligence, Springer-Verlag, pp. 113-132.

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

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

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