Automatic Extraction of Verb Phrases from Annotated Corpora: A Linguistic Evaluation for Estonian

Gaël Dias† and Heiki-Jaan Kaalep‡ and Kadri Muischnek‡

†Computer Science Department
Beira Interior University
Portugal
ddg@noe.ubi.pt

‡Department of Estonian and Finno-Ugric Linguistics
University of Tartu
Estonia
{hkaalep,kmuis}@psych.ut.ee

Keywords: Verb Phrase Extraction, Mutual Expectation, GenLocalMaxs, New Emerging Languages

Contact Author: Gaël Dias, Computer Science Department, Beira Interior University, rua Marquês d'Ávila e Bolama, 6200-053, Covilhã, Portugal, ddg@noe.ubi.pt

Under consideration for other conferences (specify)? No

Abstract

Statistically-based phrase extractors are fundamental tools for the improvement of Natural Language Processing applications designed for the new languages of the emerging countries. In this context, we will present a new architecture called SENTA (Software for the Extraction of N-ary Textual Associations) that identifies verbal phrases from lemmatized corpora. In particular, SENTA proposes a solution to the definition of ad hoc association measure thresholds and introduces a new association measure called the Mutual Expectation. In order to evaluate SENTA, an exhaustive linguistic analysis of the results has been carried out that takes advantage of a pre-existing manually created phrasal lexicon.

Automatic Extraction of Verb Phrases from Annotated Corpora: A Linguistic Evaluation for Estonian

1.1 Abstract

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 is verb phrases i.e. phrasal verbs like *to pay off* and idiomatic expressions like *to laugh one to pieces*. In this paper, we will present the SENTA system that proposes an innovative architecture that avoids the definition of global association measure thresholds and defines a new association measure that does not over-evaluate the degree of cohesion of sequences of words containing frequent fragments. Finally, we will present a case study to demonstrate a successful way of combining linguistic and statistical processing to extract Estonian phrasal verbs from a text corpus.

2 Introduction

In order to be able to analyze and synthesize 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. A special case of such common expressions is verb phrases i.e. phrasal verbs like *to pay off* and idiomatic expressions like *to laugh one to pieces*. A repository of phrasal verbs is indispensable for all the levels of linguistic analysis. One can create such a repository from existing dictionaries or other linguistic resources. However, their maintenance and upgrade often require a great deal of manual efforts that can not cope with the ever growing number of text corpora to analyze. Fortunately, language-independent computational tools have been developed in order to identify and extract multiword units from electronic text corpora such as Church and Hanks (1990), Gale (1991), Dunning (1993), Smadja (1993) and Dias *et al.* (2000). However, very few exhaustive evaluations have been carried out to test the validity of the extraction results. On one hand, most studies have focused on the extraction of compound nouns and names such as Justeson and Katz (1993) and Daille (1995). However, statistical extractors identify a great deal of linguistic phenomena. In particular, Dias *et al.* (2000) have shown that compound determinants, verb phrases, and adverbial, prepositional as well as conjunctive locutions are likely to be extracted. On the other hand, most evaluations have been carried out over English and do not test the ability of the statistical tools to extend to new languages. Finally, validated linguistic resources such as dictionaries of multiword units are an important source of knowledge that can be used to test the performance of the acquisition process. The best way to evaluate a computational tool is by using it! This is exactly the way that SENTA (Software for the Extraction of N-ary Textual Associations) has been evaluated i.e. by using it to extract phrasal verbs. This initiative came from a group of linguists that wanted to check how well a pre-existing database of phrasal verbs and idiomatic verbal expressions for Estonian, built on the basis of human-oriented dictionaries, would reflect the actual usage of complex verbs in real texts. The evaluation procedure is simple: run the statistical extractor SENTA, find expressions among multiword unit candidates, compare the results with the existing database, and add new validated information to the database. In this paper, we will first present the SENTA system that proposes an innovative architecture that avoids the definition of global association measure thresholds and defines a new association measure that does not over-evaluate the degree of cohesion of sequences of words containing frequent fragments. Then, we will present a case study to demonstrate a successful way of combining linguistic and statistical processing to extract Estonian phrasal verbs from a text corpus.

3 Data Preparation

According to Justeson (1993), the more a sequence of words is fixed, that is the less it accepts morphological and syntactical transformations, the more this sequence is likely to be a multiword lexical unit. In particular, this assumption has lead researchers to work on unannotated corpora. However, Estonian is a highly flective language with a free word order. As a consequence, verbs and nouns may occur in the input text corpus in a variety of graphical forms. Thus, in order to capture statistical regularities, the lemmatization of the input text corpus is necessary and Justeson's assumption does not stand (at least for verb phrases). In parallel, Smadja (1993) highlights that there is strong lexicographic evidence that most lexical relations associate words separated by at most five other words. Therefore, multiword lexical units may be represented as specific contiguous or non-contiguous *n*-grams in a window of ten words long (i.e. five to the left of the pivot word and five on its right hand side). We will see further in this article that non-contiguous *n*-grams are particularly important for the acquisition process as Estonian syntax has strongly been influenced by German: the usage of phrasal verbs in Estonian is often viewed as being similar to German making great use of distant word associations. So, the first step of the system is to build all contiguous and non-contiguous *n*-grams from the lemmatized input text. As an example, if sentence (1) is the current input and w_1 =*Maastricht* is the pivot word, one non-contiguous and one contiguous *2*-grams containing w_1 are shown in Table 1.

(1) "*After difficult negotiations, the Maastricht Treaty has been modified by all the State members.*"

Table 1: Two *2*-grams retrieved from sentence (1) containing *Maastricht*

W_1	***position$_{12}$***[1]	W_2
Maastricht	-3	*Negotiations*
Maastricht	+1	*Treaty*

Obviously, not all *n*-grams are multiword units. In order to decide whether an *n*-gram is a multiword unit or not, many researchers have proposed to define the degree of cohesion between words using statistical measures.

As a consequence, the higher the degree of cohesion is, the more relevant the *n*-gram is. In the next section, we propose a new association measure called the Mutual Expectation. In particular, the Mutual Expectation evidences two important characteristics. First, it is a normalized measure. Second, it does not under-evaluate the degree of cohesion of *n*-grams containing frequent words.

4 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 new mathematical model to describe the degree of cohesiveness that stands between the words contained in an *n*-gram. The association measures presented so far in the literature (cf. Church (1990), Gale (1991), Smadja (1993), Dunning (1993), Smadja (1996)) are not satisfactory as they only evaluate the degree of cohesion between two sub-groups of an *n*-gram. Furthermore, as they rely too much on the marginal probabilities of the word occurrences, they miscalculate the cohesiveness values when the *n*-grams contain frequent fragments (Dias *et al.* 2000). In order to overcome both problems, we present the Mutual Expectation measure based on the Normalized Expectation.

4.1 Normalized Expectation

We define the normalized expectation 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 [*Council +1 of +2 Ministers*] must take into account the expectation of occurring *Ministers* after *Council of*, but also the expectation of the preposition *of* linking together *Council* and *Ministers* and finally the expectation of occurring *Council* before *of Ministers*. This situation is graphically illustrated in Table 3 where one possible expectation corresponds to one respective row.

[1] In table 1, *position$_{12}$* is the signed distance between w_1 and w_2 . The sign "+" ("-") is used for words on the right (left) of w_1 .

Table 3: Example of expectations to take into account in order to evaluate the NE

Expectation to occur the word	**Knowing the gapped 3-gram**
Council	[_____ *+1 of +2 Ministers*]
Of	[*Council +1* _____ *+2 Ministers*]
Ministers	[*Council +1 of +2* _____]

The basic idea of the normalized 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. The underlying concept of the normalized expectation is based on the conditional probability defined in Equation (1).

$$p(X = x \mid Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)} \qquad (1)$$

The definition of the conditional probability can be applied in order to measure the 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. However, this definition does not accommodate the *n*-gram length factor. For example, Table 3 clearly points at three possible conditional probabilities for a *3*-gram. Naturally, an *n*-gram is associated to *n* possible conditional probabilities. As a consequence, it is clear that the conditional probability definition needs to be normalized in order to take into account all the conditional probabilities involved by an *n*-gram. One way to solve the normalization problem is by introducing the Fair Point of Expectation (FPE). In order to perform the normalization process, it is convenient to evaluate the gravity center of the denominators of all the possible conditional probabilities thus defining an average event called the FPE. Basically, it is the arithmetic mean of the *n* joint probabilities[2] of the (*n*-1)-grams contained in an *n*-gram. In other words, the FPE is defined as the average point of expectation embodying all the particular points of expectation, thus reducing the *n* particular points of expectation to just one average point. The FPE for an *n*-gram is defined in Equation (2)

$$FPE([\mathrm{w}_1\, \mathrm{p}_{12}\, \mathrm{w}_2 \ldots \mathrm{p}_{1i}\ \mathrm{w}_i \ldots \mathrm{p}_{1n}\ \mathrm{w}_n]) = \frac{1}{n}\left(\begin{array}{l} p([\mathrm{w}_2 \ldots \mathrm{p}_{2i}\ \mathrm{w}_i \ldots \mathrm{p}_{2n}\, \mathrm{w}_n]) + \\ \sum_{i=2}^{n} p\left(\left[\mathrm{w}_1 \ldots \hat{\mathrm{p}}_{1i}\ \hat{\mathrm{w}}_i \ldots \mathrm{p}_{1n}\, \mathrm{w}_n\right]\right) \end{array} \right) \qquad (2)$$

where the "^" corresponds to a convention frequently used in Algebra that consists in writing a "^" on the top of the omitted term of a given succession indexed from 1 to n. Hence, the normalization of the conditional probability is realized by the introduction of the fair point of expectation into the general definition of the conditional probability. The symmetric resulting measure is called the normalized expectation and is proposed as a "fair" conditional probability. It is defined in Equation (3).

$$NE([\mathrm{w}_1 \ldots \mathrm{p}_{1i}\, \mathrm{w}_i \ldots \mathrm{p}_{1n}\, \mathrm{w}_n]) = \frac{p([\mathrm{w}_1 \ldots \mathrm{p}_{1i}\, \mathrm{w}_i \ldots \mathrm{p}_{1n}\, \mathrm{w}_n])}{FPE([\mathrm{w}_1 \ldots \mathrm{p}_{1i}\, \mathrm{w}_i \ldots \mathrm{p}_{1n}\, \mathrm{w}_n])} \qquad (3)$$

4.2 Mutual Expectation

Daille (1995) shows that one effective criterion for multiword lexical unit identification is simple frequency. From this assumption, we pose that between two *n*-grams with the same normalized 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.

$$\begin{aligned} ME([w_1 \ldots p_{1i}\, w_i \ldots\ p_{1n}\, w_n]) &= f([w_1 \ldots p_{1i}\, w_i \ldots\ p_{1n}\, w_n]) \\ &\times NE([w_1 \ldots p_{1i}\, w_i \ldots\ p_{1n}\, w_n]) \end{aligned} \qquad (4)$$

So, the Mutual Expectation between *n* words is defined in Equation (4) based on the normalized expectation and the relative frequency. From the set of all valued *n*-grams, it is then necessary to extract the pertinent items. For that purpose, most of the studies have proposed to define association measure thresholds that divide the search space into two subsets: one for the pertinent n-grams and one for the other ones. However, this coarse grain methodology presents many drawbacks. In order to overcome this situation, we present a new algorithm based on the analysis of local maxima: the GenLocalMaxs.

[2] In the case of *n*=2, the FPE is the arithmetic mean of the marginal probabilities.

5 The GenLocalMaxs Algorithm

Being the association measure value associated to each *n*-gram, the only feature available to the system in order to extract multiword unit candidates, most of the approaches proposed in the literature have based their selection process on association measure thresholds (cf. Church (1990), Daille (1995), Smadja (1996) and Shimohata (1997)). This is defined by the underlying concept that there exits a limit value of the association measure that allows to decide whether an *n*-gram is a multiword lexical unit or not. However, these thresholds can only be justified experimentally and so are prone to error. Moreover, the association measures tend to favor certain properties of the multiword lexical units and as a consequence, the coarse grain threshold methodology may unjustifiably reject potential expressions in the set of all valued *n*-grams. Finally, the thresholds may vary with the type, the size and the language of the document and vary obviously with the association measure. The GenLocalMaxs algorithm, based on local maxima association measure values, proposes a more robust, flexible and fine-tuned approach for the election of multiword lexical units. The GenLocalMaxs elects the multiword units from the set of all the cohesiveness-valued *n*-grams based on two assumptions. First, the association measures show that the more cohesive a group of words is, the higher its score[3] 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 association measure value, *val(W)*, is a local maximum. Let's define the set of the association measure values of all the (*n-1)*-grams contained in the *n*-gram *W*, by Ω_{n-1} and the set of the association measure 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 $W$=2 then
        if $val(W) > val(y)$ then $W$ is a multiword unit
else
        if $val(x) \leq val(W)$ and $val(W) > val(y)$ then W is a
        multiword unit
```

Figure 1 : The GenLocalMaxs

So, the GenLocalMaxs algorithm avoids the *ad hoc* definition of any global association measure threshold and focuses on the identification of local variations of the association measure values. This methodology overcomes the problems of reliability and portability of the previously proposed approaches. Indeed, any association measure that shares the first assumption (i.e. the more cohesive a group of words is, the higher its score will be) can be tested on this algorithm. Finally, we propose in the next section an exhaustive evaluation of SENTA for the specific extraction of verb phrases over a text corpus written in Estonian, a language from the new emerging countries.

6 An Exhaustive Evaluation

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 advantages of the specificities of the language. Indeed, it is likely to find inflectional endings that may weaken the results of the extraction. Eliminating

[3] The conditional entropy measure is one of the exceptions.

them, by using a lemmatizer, would give statistical software better grounds for finding recurring patterns. Consider, for example, expressions like *saalomonlik otsus* (*Salomon's decision*) where both components may inflect freely. 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 abovementioned 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: lemmatize only some words (the ones that inflect freely in the expressions), and do not lemmatize others.

SENTA has been evaluated for the extraction of verb phrases over a 500,000 words sub-corpus extracted from the Corpus of Written Estonian of the 20th Century that is available at `http://www.cl.ut.ee/en/corpusb/1990s.html`. For this specific task, we have also used a pre-existing database of Estonian phrasal verbs containing 10816 entries that has been built from a set of Estonian resources aimed at a human reader[4]. Thereafter, we propose the global schema of our experiment. For the purpose of the explanation, we will call it SENVA (Software for the Extraction of N-ary Verbal Associations).

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 run the GenLocalMaxs. Based on these, manually check the extracted phrases.

In order to evaluate the set of contiguous and non-contiguous extracted verb phrases, this experiment has been realized four times, each time setting a different limit (0 to 3) to the number of words that may intervene the words that belong to a phrase. For the global evaluation, we combined the results of the four experiments and compared them against our existing database. For that purpose, we manually checked all the extracted phrases that were not in the database, and decided whether they should be added or not. 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 that a linguist would rather not present in the database although they may show usefulness for specific NLP applications such as Machine Translation and Information Retrieval. These results of precision are quite low. In fact, we expected better figures such as the ones that SENTA had shown when applied to all kinds of linguistic phenomena embodied by the multiword units (Dias *et al.* 2000). However, we should point at the fact that the extraction of verb phrases is a much more difficult task than the one for noun phrases. Indeed, verb phrases are less regular than nominal associations being more flexible in terms of word organization. As a consequence, statistical approaches evidence difficulties in putting forward verb usage

[4] In particular, this lexicon has been built from the following resources: the Explanatory Dictionary of Estonian (EKSS), Saareste (1979), Hasselblatt (1990), Õim (1991)’s Dictionary of phrases, Õim (1993)’s Dictionary of synonyms and the Filosoft thesaurus (http://ee.www.ee/Tesa/).

regularities. These low results clearly show why very few studies have been proposed for the evaluation of statistical extractors when applied to verb phrase identification. Indeed, SENTA has proved to behave as well as most of the important extractors in the context of multiword unit extraction (Dias *et al.* 2000) and so, these figures are likely to be the same as most of the statistical tools. Nevertheless, these results must be deeply analyzed in order to draw final conclusions. In fact, 1629 of the 2,500 were expressions that the database already contained and SENVA found 865 more phrases that should be included. Table 4 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 4: Some phrasal verbs in the database and/or text corpus

6.1.1.1 Phrase	**In the DB**	**Found by SENVA**
biellu astuma (to marry)	+	-
abiellu heitma (to marry)	+	-
abielu rikkuma (to commit adultery)	+	-
abielu sõlmima (to contract a marriage)	+	-
abielu lahutama (to divorce)	-	+
allkirja andma (to give one’s signature)	-	+
andeks andma (to forgive)	+	+
andeks paluma (to apologize)	+	-
andeks saama (to obtain forgiveness)	-	+
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)	-	+

These figures give us an estimation of the quality of the database. Indeed, we can see that out of the 2,500 phrasal verbs that were extracted from the corpus, only 65% were represented in the database. Without SENVA, we would not have been able to find the missing 35%. This is an important point that can be credited to statistical extractors. They are likely to find a great proportion of unknown phrasal verbs. Moreover, evaluating the results of the acquisition process also implies to answer the following question: how sure can we be that SENVA really found all the phrasal verbs that are in the corpus (recall), and that it did not report about phrasal verbs that are not used as such in the corpus? For estimating this, we made an experiment with 500 randomly selected phrasal verbs of our database. By checking the corpus manually, we found that 131 out of the 500 could be found in the corpus. In principle, SENVA can only find phrases that occur at least twice in the corpus. In this context, the number of such phrases was 71. So, we made 4 experiments with SENVA, where we defined different numbers of words that could possibly occur between the words of a phrase: 0[5], 1, 2 and 3. The number of correct phrases that SENVA found is shown in Table 5.

Table 5: Number of Extracted Phrases

Distance	0	1	2	3	Combined
# of Phrases	45	46	50	52	57

For a distance of 3 words, among the 19 phrases that SENVA did not find, 12 occurred in the corpus twice, but 5 were rather frequent. They are illustrated in Table 6.

Although SENVA may have failed to find phrases that are in our database, it is interesting to notice that it often found phrases that contain the ones in our database. For example, 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)*. In this specific case, SENVA has pointed to an error in our database. Indeed, 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

[5] This stands for contiguous verb phrases.

be discarded as a phrasal verb and replaced with a finer-grained unit.

Table 6: Missed Phrases

Phrase	# of co-occurrence s	# of phrases[6]
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)	21	19
ära maksma (to pay off)	12	9

As a summary of this exhaustive evaluation, if 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 would find 57/71=80% of those that occur more than once and close to 99% of those that occur more that 3 times which evidences a very high recall rate thus balancing the lower precision results.

7 Conclusion

In this paper, we presented a new statistical tool called SENTA (Software for the Extraction of N-ary Textual Associations) that introduces two important concepts that overcome important drawbacks of existing extractors: the Mutual Expectation and the GenLocalMaxs algorithm. In order to evaluate this new architecture, we performed an experiment over a 500,000 words Estonian corpus thus taking advantage of our extractor's adaptability to new languages. Thus, we proposed an evaluation for the difficult task of verb phrase extraction based on a highly flective language, Estonian. The results showed that although precision is surprisingly low, recall overtook all our expectations. Thus, the manually checked extracted phrasal verbs were directly used to update (i.e. add and correct) an existing database of multiword units for Estonian.

8 References

Church K. (1990) *Word Association Norms Mutual Information and Lexicography*, Computational Linguistics, 16/1, pp. 23--29.

Daille B. (1995) *Study and Implementation of Combined Techniques for Automatic Extraction of Terminology*, The balancing act combining symbolic and statistical approaches to language, 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*, Journal Traitement Automatique des Langues, Vol 41:2, Christian Jacquemin (ed.). Paris, France, 2000, pp 447-473.

Dias, G., Guilloré, Lopes, J.G.P. (2000) *Mining Textual Associations in Text Corpora*, Workshop on Text Mining of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, MA, USA, pp. 92-95.

Dunning T. (1993) *Accurate Methods for the Statistics of Surprise and Coincidence*, Association for Computational Linguistics, 19/1.

Gale, W. (1991) *Concordances for Parallel Texts*, In "Proceedings of the Seventh Annual Conference of the UW Center for the New OED and Text Research, Using Corpora" Oxford, England.

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

Justeson J. (1993) *Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text*, IBM Research Report, RC 18906 (82591) 5/18/93.

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

Õim, A. (1991) *Sünonüümisõnastik (Dictionary of synonyms)*, Tallinn, Estonia.

Saareste, A (1979) *Eesti keele mõistelise sõnaraamatu indeks (Index of the thesaurus of Estonian)*, Finsk-ugriska institutionen, Uppsala.

Shimohata S. (1997) Retrieving Collocations by Co-occurrences and Word Order Constraints, In "Proceedings of ACL-EACL'97", pp. 476—481.

[6] We differentiate a co-occurrence from a phrase by the fact that a co-occurrence may not be used as a phrase in a given context.

Smadja F. (1993) *Retrieving Collocations From Text: XTRACT*, Computational Linguistics, 19/1, pp. 143—177.

Smadja F. (1993) Translating Collocations for Bilingual Lexicons: A Statistical Approach, Association for Computational Linguistics, 22/1.