Methods for the Extraction of Hungarian Multi-Word Lexemes

Balázs Kis, Begoña Villada Moirón◇, Tamás Bíró◇,*
Gosse Bouma◇, Gábor Pohl, Gábor Ugray*, John Nerbonne◇*

◇ Rijksuniversiteit Groningen * MorphoLogic, Budapest

Abstract

This paper describes an experiment on extracting Hungarian multi-word lexemes from a corpus, using statistical methods. Corpus preparation—the addition of POS tags and stems—was done automatically. From the corpus, ⟨*verb+noun+casemark*⟩ patterns were extracted as collocation candidates. Evaluation shows that the statistical methods used by Villada Moirón (2004a) to identify Dutch V + PP collocations, can also be applied to the Hungarian data. Some collocation types (such as verbal arguments) require special extraction methods, as explained in the evaluation section. Finally, we suggest that the extraction process can be further improved by a blend of statistical techniques with rule-based and dictionary-based methods.

1 Introduction

This paper presents partial results of a Hungarian–Dutch research project on finding and processing multi-word lexemes. The main research goal is to provide an efficient method to extract multi-word lexemes from Hungarian corpora, in order to have

1. better linguistic coverage for the Hungarian language,
2. a more efficient way to improve lexicons,
3. quality improvement in existing Hungarian linguistic applications.

The method is built on the basis of on-going research on identification of Dutch collocations by Villada and Bouma (2002) and Villada Moirón (2004a) and it uses the *ngram* statistical package (Pedersen and Banerjee 2003). So far, a corpus-based statistical method has been implemented for finding Hungarian multi-word lexemes. This presents a special challenge to the methods themselves as they were developed for a language with a very different morphology and syntax. Although there are considerable results in the field of the automatic identification of collocations and multi-word lexemes in English (Kilgarrif and Tugwell 2001), German (Kermes and Heid 2003), Dutch, etc., no previous research has been carried out for Hungarian at this level of complexity.

After providing our working definition of multi-word lexemes, we describe the tools and the statistical method used for multi-word lexeme extraction in section 2. Section 3 reports on our evaluation results and discusses a few observations. Finally, section 4 proposes improvements and further research, while we summarize our conclusions in section 5.

1.1 The working definition of multi-word lexemes

Multi-word lexemes (MWLs) are potentially a very broad concept. For a unified and coherent definition, we may start from the definition of a *lexeme* itself:

”A lexeme is the basic abstract unit of the lexicon (. . .)” (*Routledge Dictionary of Language and Linguistics* 1996).

This definition does not make assumptions about the actual wording of a lexeme, i.e. it can be either a single-word or a multi-word expression. Multi-word lexemes become a problem with natural language processing because single-word units have clear boundaries in electronic text, while multi-word ones have not.

A concept represented by a lexeme in different languages can be broken up into words in different ways. In European languages, isolation (English and Romance languages in some cases) and extreme compounding (Dutch, German, Hungarian etc.) also exist. Example (1) is taken from technical terminology.

(1)	direction assistée	(French)
	stuurbekrachtiging	(Dutch)
	szervokormány	(Hungarian)
	power steering	(English)

The above definition of a lexeme as a ’basic unit’ may seem vague; we could also use some more specific synonyms here: a lexeme is not necessarily a *minimal unit* of language, neither semantically, nor in terms of syntax. It is more appropriate to label them as *integral units*, which are not required to be *atomic* units at the same time. A lexicographer has an internalized notion of a lexeme when he chooses an expression to be included in a dictionary as a headword. In several cases, dictionary headwords are not atomic units (in terms of morphology and syntax), but they are considered as integral semantic units.

Since our goal is to apply computational tools to detect multi-word lexemes in running texts, we need a working definition specifying clear features of MWLs that can be used directly in the computational model.

Given the theoretical definition, we are looking for pieces of text forming a single semantic unit. Computational modeling of semantics is relatively difficult. Therefore, we may wish to identify more specific features of MWLs.

One such feature is (semantic) non-compositionality where the distinct sense of a lexical unit within a multiword lexeme cannot be separated, i.e. the meaning of the unit is different from the combination of the meaning of the parts. This type of non-compositionality is difficult to model—because it still requires semantics—and several multi-word lexemes do not even display this feature. If we accept that a MWL might have a non-atomic meaning too (when the MWL is a lexical, morphosyntactic or syntactic compound, having a composite meaning, but referring to one specific thing, like in the case of ‘power steering’), there remains one clear feature that an MWL can display, namely, *fixedness* (Moon 1998).

A fixed expression—chosen as a model for MWLs—is a collocation displaying syntactic irregularity and/or semantic non-compositionality. Let us emphasize that these characteristics—used as working criteria for differentiating between multi-word lexemes and ordinary collocations—do not imply syntactic non-variability, i.e. the definition still allows for internal modification of the same collocation. Here, we consider *collocations* as two or more co-occurring unigrams, without any assumptions about their dependence or independence. Let us also note that MWLs form only a subset of fixed expressions because there are fixed expressions such as proverbs and idioms that do not fulfill the original criterion of being a single semantic unit.

Co-occurrence and syntactic irregularity are surface characteristics relatively well recognizable by formal methods. Our hypothesis is that implementing these in a computational model provides for accurate approximation of the 'ideal' set of multiword lexemes. Within the scope of our experiment, the working model assumes that parts—component unigrams—of multi-word lexemes (e.g. 'take into account') combine with a better-than-chance frequency, i.e. it is more probable for them to occur together than we would expect on the basis of their individual frequencies.

By our working definition, the necessary but not always sufficient defining criteria of an MWL is some degree of lexical fixedness; syntactic variability is allowed. However, for some purposes, we still need to look for semantic non-compositionality:

1. In machine translation, we have to determine if a source-language expression may be translated as a combination – or different wording must be used.

2. In computational terminology, we have to determine if a given expression belongs to the semantic domain in question.

In section 4.3, where we relate our work to the results of other researchers, aspects of computational terminology are emphasized to a great extent. Although terminology is a somewhat narrower concept than multi-word lexemes, published methods of terminology extraction and enrichment are very similar to those described here.

1.2 Cross-language and cross-platform investigation

The resource preparation process is not part of the *ngram* package. Therefore, new dataset extraction procedures were required to provide suitable input. To this end, we have developed a new collocation candidate extraction package primarily for use with Hungarian texts, so that we can compare our results to those in other languages (Kis, Villada, Bouma, Bíró, Nerbonne, Ugray and Pohl 2004).

This step was taken on the basis of the assumption that a computational method is proved stronger if it may be transferred to other languages and/or computational platforms and remains successful, i.e. continues to produce useful results.

2 Statistical investigation of Hungarian collocations

The extraction process for the Hungarian collocations has been designed to provide results compatible with those achieved earlier by Villada Moirón (2004b) and Villada Moirón (2004a). This applies both to the types of collocations extracted, and the way the intermediate results are formatted.

2.1 Extraction tools

Collocation candidates are identified using a Hungarian parser named HumorESK.[1] This parser is capable of identifying named entities, NPs, VPs, and most sentence structures, and it provides a fairly deep parse. This package provides for intelligent collocation searching as one is able to specify the types of the components of the collocations and a window size in terms of terminal symbols, within which the co-occurrence must appear. With this system, a terminal symbol is either a word or a punctuation mark (with the exception of periods at the end of abbreviations or decimal points—in Hungarian, commas—within numbers).

Parsing accuracy of the HumorESK parser is currently being measured. As of now, we have estimates that indicate a 70 to 98 percent recall in finding NP heads depending on the type of the text parsed. Precision is difficult to estimate because on the one hand, the parser identifies all partial NPs, on the other hand, it performs a heuristic filtering by applying highest-level NP rules looking for a longest-match. This causes occasional problems with detecting NP boundaries. These are partially corrected by applying higher level parsing, thus some of these problematic parses are discarded. In Hungarian, the detection of NP boundaries is closely related to finding the NP head because the head is almost always the last word in the NP.

Within the project, we are using the *ngram* program,[2] developed entirely in Perl by Pedersen and Banerjee (2003). The *ngram* package was only used to apply the statistical tests and rank the ngrams (expressions) in datasets. The dataset is a collection of candidate ngrams, their observed frequency in the ⟨*verb* + *noun* + *casemark*⟩ event space, as well as the frequency of its component unigrams and partial bigrams. This very same scheme is used by Villada Moirón (2004a).

2.2 The corpus

On site, we had the recently compiled SZAK Corpus (Kis and Kis 2003) available, an English-Hungarian parallel corpus of technical texts. The SZAK Corpus has been compiled by a publishing company who have now been purposefully working on a corpus of publications. The corpus consists of a monolingual subcorpus of

[1] The name stands for High-Speed Unification-based Morphology Enriched by Syntactic Knowledge (Pr´osz´eky 1996), which indicates that it is a bottom-up parser based on lookups of finite syntactic patterns in a lexicon. In the earlier versions, the entire grammar used to be 'finitized' into a single lexicon by means of RTNs, and thus its operation was very similar to that of HuMor, MorphoLogic's morphological analyzer.

[2] Formerly known as NSP (Ngram Statistics Package), *ngram* is a SourceForge project available at `http://sourceforge.net/projects/ngram`.

original Hungarian works in the field of computing, with a size of approx. 500,000 words, and a bilingual parallel corpus of translated works in computing, with approx. one million words per language.[3] The main goal of the corpus is to facilitate research on terminology and translation studies, with text structures taken into consideration. This corpus did not have morphosyntactic annotation at the time.

During the present experiment, we used the Hungarian components of the corpus (the monolingual part and the Hungarian texts from the bi-lingual subcorpus) for the statistical tests. The Hungarian subcorpus contains approximately 1.5 million words altogether. A corpus of this size is suitable for seeking the most frequent multi-word lexemes. It is too small to be used to detect infrequent combinations.

2.3 Types of collocations extracted

In the experiment, we investigated ⟨*verb* + *noun* + *casemark*⟩ patterns as candidates, where the noun is in fact the head of an NP which has the casemark attached as a suffix. This structure has been selected in order to provide results that are to some extent comparable or compatible with those acquired by Villada (ms). We wished to find a Hungarian construction that most closely corresponds to ⟨verb + PP⟩ collocations studied earlier in the Dutch experiment.

In Hungarian, the role of prepositions is fulfilled by a case suffix at the end of an NP's head, which is almost always at the end of the NP itself:

(2) *Az utca vég-é-n*
The street end-POSSESSED-SUPERESSIVE
'At the end of the street'

The typological difference between Hungarian and Dutch makes therefore the experiment especially interesting. The Hungarian morphological analyzer separates the case suffix from the lemma, then the parser classifies the case itself based on the surface form of the suffix.

Note that Hungarian also has postpositions, but they express more unusual grammatical relations than the approximately twenty cases. Therefore, their role in Hungarian is significantly smaller than the role of prepositions in English or Dutch. This is why we have ignored them in our experiment.

2.4 The extraction method

Based on the HumorESK parser described in section 2.1, a special candidate extractor was written that takes a single meta-rule as the description of the collocation we are looking for:

VX!(lex),NP-FULL!(lex,case):5

[3] The texts are copyrighted, but the copyright proprietors have granted us the right to use them for research purposes, provided that the texts will not be re-published in their entirety. The corpus includes the full text of all electronically available publications on the backlist of the publisher, resulting in the corpus size mentioned above.

Using this metarule, the extractor program will extract trigrams consisting of

- the lemma of a verb,
- the lemma of the head of an NP, and
- the casemark of (the head of) the same NP,[4]

where the verb and the NP occur within a window of 5 terminal symbols. This window is far smaller than the one used by Villada (ms.). It was intentionally chosen – after some empirical experiments – in order to limit the otherwise significant level of noise (i.e. irrelevant co-occurrences of verbs and noun phrases), introduced partly by the lack of disambiguated POS tagging.

In addition to the surface and lexical forms, the program is capable of extracting any feature of a node in the parse tree. The need for this is obvious if we aim at extracting Hungarian entities corresponding to prepositions—that are in fact the types of case marks attached to nouns.

Without disambiguated POS tagging, morphological analysis and parsing is run in a single process. Although parsing is rather deep and the nature of both the parser and the grammar allow for rules (patterns) overriding other rules, some parsing errors can still occur due to morphological misclassification.

Most of these errors are quite obvious and directly related to the types of errors the morphological analyzer is most prone to. We have reviewed a number of morphological misclassifications, and developed a filtering mechanism which discards some of the collocation candidates. This filtering program uses metarules similar to those of the extractor. These rules are entirely heuristic and are based on morphological ambiguities where the less probable interpretation (e.g. a noun instead of a number) might have been used to build an NP, a VX, or any other node in the parse tree.

This post-filtering mechanism is in place only to limit the misclassification noise in the system, and is inactive when a Hungarian POS tagger is present.

2.5 Statistical measures applied

When selecting statistical functions to apply to our data sets, we have relied on the evaluation of the Dutch experiments (Villada, ms). We have disregarded those functions that the Dutch evaluation found less precise, and we selected two that provided the best precision with the best recall. These two functions were log likelihood (Dunning 1993) and salience (Kilgarrif and Tugwell 2001). Both measures assess the components of an ngram occurring together against each component occurring independently.

The log likelihood score of a bigram is the ratio between two likelihoods: (i) the likelihood of seeing one component of a collocation given that another is present,

[4] We cannot really talk about heads in HumorESK as it does not use a head-driven grammar. However, it specifies explicit feature inheritance: if the grammar has been written accordingly, one can treat the 'lex' feature of an NP node as the 'lex' of its head.

and (ii) the likelihood of seeing the same component of a collocation in the absence of the other. When the ratio is large, we have evidence of statistical dependence. To compute the log likelihood ratio, we use a simpler formula, namely, the log likelihood chi-square ratio G^2 (see Agresti (2002)). For a bigram (w_i, w_j), G^2 adds up the product of the observed bigram frequency O_{ij} and the logarithm result of dividing the observed frequency O_{ij} by the expected bigram frequency E_{ij}:

$$G^2 = 2 \sum_{i,j} O_{ij} \log_2 \frac{O_{ij}}{E_{ij}}$$

The salience measure is an adjustment to the mutual information test. Mutual information ($I(w_i, w_j)$) compares the probability of seeing the unigrams in an ngram together to the probability of the independent occurrence of each. The salience adjustment multiplies the mutual information score by the logarithm of the ngram's observed frequency, thus it promotes the frequent ngrams to the top ranks. Kilgarrif and Tugwell (2001) calculate it as follows:

$$\mathit{Sal}(w_i, w_j) = I(w_i, w_j) \log_2 O_{ij} = \log_2 \frac{P(w_i, w_j)}{P(w_i)P(w_j)} \log_2 O_{ij}$$

The two statistics are designed to assess the dependence between two random variables (two unigrams). Because our candidate expressions are ⟨*verb, noun, casemark*⟩ trigrams, we decided to represent each candidate trigram by two bigrams: ⟨ *verb, noun_casemark*⟩ and ⟨*verb_noun, casemark*⟩. Bearing this in mind, we first compute both statistics for each bigram. Bigrams are ranked according to the statistic so that, candidate bigrams which were assigned a large score will get a higher rank than bigrams assigned a small score. In a next step, the ranks of each partial bigram are added up. We order the candidate trigrams according to the resulting rank. The trigrams with the highest ranks are those whose association measure score was higher overall.

3 Evaluation

We have evaluated the results of ranking the candidates in detail. We looked at the results from both the log likelihood and the salience functions. Because this was the very first experiment with this procedure and these texts, we had no hand-tagged data to compare the results to. Thus we decided to manually check those candidates ranked among the first 100 either by the log likelihood or the salience functions.

Manual checking was performed by 3 independent human judges assigning binary scores: a candidate was assigned '0' if it was not a multi-word lexeme recognizable to the judges; '1' if it was either a compositional or a non-compositional expression. In a voting scheme, if a candidate received two '1' votes (score 2), it was accepted as a valid expression. Then the judges determined whether a candidate, having received a score of 2 or 3, is compositional or non-compositional. As a baseline, we used the ranking provided by the raw frequency list, i.e. ordering

Measure	Multi-word lexemes	non-compositional	transparent
Salience	82	25	57
Log likelihood	76	18	58
Raw frequency	36	6	30

Table 1: Number of multi-word lexemes among the 100 highest ranked trigrams.

Measure	Incorrect patterns
Salience	10
Log likelihood	15

Table 2: Evaluation of the influence of parse errors on the 100 highest ranked phrases.

the candidates in dataset according to the number of occurrences of each candidate in the corpus.

Table 1 shows that both the salience measure and the log likelihood test contain a substantial number of multi-word lexemes in their top 100. Both tests also outperform the baseline. The candidate rankings produced by the log likelihood and salience tests are rather similar, with a *Spearman's correlation coefficient* $\rho = 0.94$ calculated over the ranks of expressions. Nevertheless, the salience method seems more reliable as it produces less noise in the 100-best list.

A number of collocations were classified as 'real multi-word lexemes' whose meaning was entirely non-compositional, regardless of the technical domain of the corpus. The majority of the multi-word lexemes can be classified as 'transparent', i.e. semantically compositional, by common sense. However, many compositional combinations form a terminological collocation relating to the technical field of the texts in the corpus, such as the trigram ⟨kattint + parancsgomb + SUB⟩ in example (3).

(3) *kattint parancsgomb-ra*
click button SUBLATIVE
'click a button'

Thus, most of these 'transparent' collocations are indeed important from the aspect of translation, since these should be translated consistently.

3.1 Error analysis

There were some instances of noise that remained in the dataset. Some errors were ranked surprisingly high, considerably lowering the overall precision of the statistical methods (See Table 2).

Parse errors typically lead to a situation where human reviewers are unable

Collocation	Rank by salience	Rank by log-likelihood
⟨vesz + figyelem + ILL⟩	17	41
*⟨veszik + figyelem + ILL⟩	24	70
⟨vesz + igény + ILL⟩	11	31
*⟨veszik + igény + ILL⟩	40	$>$ 100

Table 3: Morphological ambiguity represented in the Hungarian collocation tests

to reconstruct a phrase corresponding to the pattern. These problematic co-occurrences may be caused by three reasons:

1. morphological misclassification or unwanted ambiguity;
2. accidental co-occurrence of a verb and an NP where the NP is not a real argument of the verb;
3. the unigrams are part of a larger fixed expression.

The second type of error can be eliminated if we use deeper parsing instead of NP chunking and lemmatizing. Thus we could determine if the V and the NP belong to the same VP subtree and list only those co-occurrences where they do.

In the following, we present two further observations that account for errors, and help in improving the tools we used.

3.2 Hungarian morphological ambiguity spotted and resolved correctly

Let us consider the Hungarian verbal form *veszik*, having two possible morphological analyses:

1. It is the 3rd person plural form of the base verb *vesz*, meaning ‘take’. In this meaning, it frequently co-occurs with the illative form of the noun *figyelem* [‘attention’], because the multiword lexeme *figyelembe vesz* (plural: *figyelembe veszik*) mean ‘take something into account’.
2. It is also the 3rd person singular form of the base verb *veszik* [‘to be lost’]. Obviously, the latter verb cannot be combined with the illative case of *figyelem* (the result would be nonsense such as ‘to be lost in attention’).

However, the form *veszik* within a corpus occurrence of the combination *figyelembe veszik* (‘they take into account’) will be automatically parsed in both ways by the morphological analyzer. Consequently, this token increases the number of ⟨vesz, figyelem, ILL⟩ trigrams found, as well as that of ⟨veszik, figyelem, ILL⟩. Obviously, cases of the singular form *figyelembe vesz* count only as the first one.

Table 3 shows the ranks of these two collocations, both by the log-likelihood and the salience measures. The table shows another frequent collocation (*igénybe* *vesz*) of the same base verb *vesz* with the illative of the noun *igény* [‘claim, demand’] with a non-compositional meaning ‘make use of something’. This collocation, quite naturally, displays the same ambiguity inherent in the verb form.

Note that the salience function ranked both incorrect interpretations rather highly—while ranking the correct ones even higher. This can be attributed to the fact that (a) the salience measure is adjusted using the frequency of the collocation, and (b) these collocations (using the 3rd plural form, which displays the ambiguity in question) occur fairly often in the technical corpus we have been investigating.

However, when looking at the measures together, we can see that the correct interpretation was ranked considerably higher.

Using this information, we can recognize different analyzes of ambiguous words, where the collocations of different interpretations are ranked differently. Here the more highly ranked interpretation is very likely to be part of a multiword lexeme, and it is possible that ambiguous words can be very efficiently disambiguated when they occur in these collocations.[5] (This observation, however, requires further investigation.)

3.3 Casemarked NP’s as verbal affixes

Similarly to Dutch and German, Hungarian has verbal affixes that form one word with the verb only when the affix immediately precedes the verb. As the following examples including the verb *bemegy* (‘go in, enter’) show, different syntactic factors (the presence of an auxiliary, focus, negation,...) may alter the position of the verbal affix. (The hyphens in (4) were inserted for clarity, and do not appear in written texts.)

(4) a. *A gyermek be-megy a házba.*
The child in-go the house-ILLATIVE
‘The child enters the house.’

b. *A gyermek nem megy be a házba.*
The child not go in the house-ILLATIVE
‘The child does not enter the house.’

c. *A gyermek be fog menni a házba.*
The child in FUTURE go-INF the house-ILLATIVE
‘The child will enter the house.’

Furthermore, some nouns, including their case ending, are on the way to becoming similar verbal affixes. Their syntactic behavior is similar, and orthographic tradition writes some of them joint to the verb whenever they immediately precede it. For instance, the noun *lét* (‘being, existence’) in the sublative case (*létre*) and the verb *jön* (‘to come’) have created the complex verb *létrejön* (‘come into being, be

[5] Note that we needed non-disambiguated corpus annotation to make this observation.

Collocation			Salience	Log-likelihood
⟨hoz + lét + SUB⟩	*l´etrehoz*	‘create’	1	1
⟨vesz + ész + SUB⟩	*´eszrevesz*	‘notice’	8	12
⟨jön + lét + SUB⟩	*l´etrej¨on*	‘come into being’	13	17

Table 4: Affixed verbs written as separate words in special cases only

established’):

(5) a. *A szervezet létre-jön*
The institution being-SUBLATIVE-come
‘The institution gets established.’

b. *A szervezet nem jön létre*
The institution not come being-SUBLATIVE
‘The institution does not get established.’

c. *A szervezet létre fog jönni*
The institution being-SUBLATIVE FUTURE come-INF
‘The institution will get established.’

When the latter structures occur as separate words, they are understood together very strongly as a collocation. These are obviously the strongest candidates for multi-word lexemes (see Table 4); however, they are already included in almost all Hungarian dictionaries as single verbs.

When written separately, the dataset extraction process spotted these occurrences as ⟨*verb*+*noun*+*casemark*⟩ collocations, and the statistics ranked these instances very high. The ranks on the table might be misleading: it is more convincing if we point out that these collocations are in fact the first occurrences of each verb in the ranked dataset, according to both measures.

Whenever the noun behaving as a verbal affix was spelled in one word with the verb, then the dataset extraction process saw it as a (complex) verb to which an additional NP head was added to form the ⟨*verb* + *noun* + *casemark*⟩ trigram. Unlike the case when the verbal affix was seen as an argument of the verb, these latter trigrams are indeed far from being multi-word lexemes. In order to assess the statistical techniques, we have, therefore, compared the rank of the trigram ⟨hoz + lét + SUB⟩ (‘bring into being’) to the rank of trigrams like ⟨létrehoz + fájl + ACC⟩ (‘create a file’). The difference in ranking is highly significant: the latter type, even when possibly forming a terminological collocation, do not appear within the 250 highest ranked candidates.

4 Suggestions for improvement

4.1 Improving the statistical methods

In order to improve the precision and reliability of the statistical extraction process, three issues need to be addressed:

1. The accuracy of the morphological analysis of the corpus text should be increased either by means of introducing a (weakly) disambiguating POS tagger, or through improving the Hungarian morphological analyzer in general.

2. Larger corpora and corpora on other domains should be used instead of, or in addition to the current technical corpus. The acquisition of larger corpora, compilation of a new corpus and enrichment of the technical corpus are in progress.

3. The Hungarian parser should be used more efficiently. The present extraction scheme uses the verbal (VX) and NP nodes independently; all it checks for is their co-occurrence within a specified window. We expect to improve this by ensuring that the co-occurring nodes, i.e. the components of the collocation, are in fact children of the same VP node. We suppose that some of the 'false argument' errors can be eliminated this way.

It is also possible to improve dataset extraction through detailed corrections. This includes the proper treatment of verbal affixes, which are at present discarded by the parser at a lower level. Furthermore, nominal postpositions—rare in Hungarian, and therefore presently ignored—could be processed as further case categories.

Other types of collocations should also be investigated. For example, a ⟨V + casemark⟩ dataset could be derived from the ⟨*verb* + *noun* + *casemark*⟩ collocations. This would provide important information on Hungarian verbal argument structures. However, this can produce sensible results if and only if the relatively free Hungarian argument structure is taken into account. Thus the verb and any of its arguments can form a candidate, regardless of their relative position as long as the surface order is concerned.

4.2 Introducing further methods

Statistical measures calculated from monolingual corpora are only one technique for identifying multiword lexemes. The extraction of multiword lexemes could potentially be implemented as a blend of different methods, of which statistical investigation is only one, though not unimportant.

As mentioned in Section 2.1, statistical methods are able to detect fixed expressions, without regard to the eventual semantic non-compositionality—which may indeed need to be detected in some cases, where multi-word expressions may not differ in combinatorial frequency.

There are fields of collocation research where both rule-based and statistical methods are applied. Real strength lies in the combination of the two. As we have access to many bilingual dictionaries and are working to develop various translation (support) tools, it is an obvious step to investigate collocations through their translations. If a collocation has a non-compositional translation in another language, chances are that its meaning is not compositional either.

Our most ambitious—non-statistical—idea of approaching multi-word lexemes is a contrastive method that aims at examining collocations through their translations in another language. This approach is still being worked on. However, the basic idea is a very practical one: it proposes using a blend of several methods, including corpus-based statistical investigation, dictionary-based lexicon enrichment and contrastive tests based on phrase-level alignment. We would require a parallel corpus aligned at the sentence level, and a high-quality bilingual dictionary. If there is a given type of collocation in the text of one language (the one under investigation), such as an NP or a ⟨V + NP + casemark⟩ pattern, and there is at least one word in it whose obvious (dictionary-based) translations cannot be found in the alignment pair, it will be a good candidate for further testing.

4.3 Relation to similar fields

There are research fields similar to the generalized research on multiword lexemes that deserve attention. One such field is computational terminology (Castellví, Bagot and Palatresi (2001), Jacquemin (2001)), which employs more or less the same methods, namely,

1. statistical collocation testing,
2. collocation extraction by means of parsing,
3. dynamic methods that start from a reference glossary and use either statistical or heuristic methods to enrich it.

Another closely related field is named entity recognition, whose methods have good and simple examples of learning extraction rules. The Hungarian parser in fact employs a robust (still heuristic) named entity recognizer implemented as a subset of its grammar.

5 Conclusion

We have investigated Hungarian multi-word lexemes. We applied a reliable statistical scheme (used earlier by Villada (ms.) and Villada Moirón (2004a)) to a new and typologically different language.

The evaluation of the initial experiments shows that the same statistical methods can be applied to Hungarian corpora, however, some collocation types (such as verbal arguments) require special extraction methods.

This project attempts to improve on existing methods not by inventing entirely new extraction schemes; the emphasis is rather put on constructing an efficient blend of existing—heuristic, statistical or dictionary-based—methods. These methods can be used as cascaded modules, or in a parallel manner, 'voting' on each candidate.

The ongoing work is directed towards improving the dataset extraction procedures for the statistical methods, and, at the same time, developing a dictionary-based methods for testing candidates through their translations.

Acknowledgments

This work has been carried out in parallel with similar work on Dutch corpora by a joint Dutch-Hungarian research group supported by NWO-OTKA under grant number 048.011.040.

References

Agresti, A.(2002), *Categorical Data Analysis*, John Wiley and Sons, New York.

Castellví, M. T. C., Bagot, R. E. and Palatresi, J.(2001), Automatic term detection: A review of current systems, *in* D. Bourigault, C. Jacquemin and M.-C. L'Homme (eds), *Recent Advances in Computational Terminology*, John Benjamins, Amsterdam–Philadelphia, pp. 53–88.

Dunning, T.(1993), Accurate methods for the statistics of surprise and coincidence, *Computational linguistics* **19**(1), 61—74.

Jacquemin, C.(2001), *Spotting and Discovering Terms through Natural Language Processing*, MIT Press, Cambridge (Mass.).

Kermes, H. and Heid, U.(2003), Using chunked corpora for the acquisition of collocations and idiomatic expressions, *Papers in Computational Lexicography: Proceedings of COMPLEX 2003*, Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, pp. 37–46.

Kilgarrif, A. and Tugwell, D.(2001), Word sketch: extraction and display of significant collocations for lexicography, *Proceedings of the 39th ACL and 10th EACL- workshop 'Collocation: computational extraction, analysis and explotation'*, Toulouse, pp. 32–38.

Kis, A. and Kis, B.(2003), A prescriptive corpus-based technical dictionary. development of a multi-purpose technical dictionary, *Papers in Computational Lexicography: Proceedings of COMPLEX 2003*, Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, pp. 47–56.

Kis, B., Villada, B., Bouma, G., Bíró, T., Nerbonne, J., Ugray, G. and Pohl, G.(2004), A new approach to the corpus-based statistical investigation of hungarian multi-word lexemes, *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC) 2004*, Lisbon, Portugal, pp. 1677-80.

Moon, R.(1998), *Fixed Expressions and Idioms in English - A Corpus-Based Approach*, Oxford University Press, Oxford, UK.

Pedersen, T. and Banerjee, S.(2003), The design, implementation and use of the ngram statistics package., *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, Mexico, pp. 153–158.

Prószéky, G.(1996), Syntax as meta-morphology, *Proceedings of COLING-96*, Copenhagen, Denmark, pp. Vol.2, 1123–1126.

Routledge Dictionary of Language and Linguistics(1996), Routledge.

Villada, B. and Bouma, G.(2002), A corpus-based approach to the acquisition of collocational prepositional phrases, *Proceedings of EURALEX 2002*, Center for Sprokteknologi, Copenhagen, Denmark, pp. 153–158.

Villada Moirón, M. B.(2004a), Discarding noise in an automatically acquired lexicon of support verb constructions, *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC) 2004*, Lisbon, Portugal, pp. 1859-62.

Villada Moirón, M. B.(2004b), Acquisition of Dutch support verb collocations: a model comparison. ms. Groningen University URL: `http://www.let.rug.nl/~begona/papers/svcmodels.ps`.