A New Approach to the Corpus-based Statistical Investigation of Hungarian Multi-word Lexemes

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

* Rijksuniversiteit Groningen, The Netherlands; ** MorphoLogic, Budapest, Hungary
Postbus 716. 9700 AS Groningen; Orbánhegyi út 5., H-1126 Budapest
* {villada,gosse,birot,nerbonne}@let.rug.nl; ** {kis,ugray,pohl}@morphologic.hu

Abstract

We apply statistical methods to perform automatic extraction of Hungarian collocations from corpora. Due to the complexity of Hungarian morphology, a complex resource preparation tool chain has been developed. This tool chain implements a reusable and, in principle, language independent framework. In the first part, the paper describes the tool chain itself, then, in the second part, an experiment using this framework. The experiment deals with the extraction of <verb+noun+casemark> patterns from the corpus as collocation candidates, in order to compare results to an experiment on Dutch V + PP patterns (Villada, 2004). Statistical processing on this dataset provided interesting observations, briefly explained in the evaluation section.

We conclude by providing a summary of further steps required to improve the extraction process. This is not restricted to improvements in the resource preparation for statistical processing, but a proposal to use nonstatistical means as well, thus acquiring an efficient blend of different methods.

1. Introduction

We describe the development of the necessary lexical resources and tools in order to apply statistical corpus-based methods to perform automatic extraction of Hungarian multi-word lexemes.

A corpus-based investigation of Hungarian multi-word lexemes is essential for the further development of a Hungarian parser, for the purpose of information extraction and machine translation/computer-aided translation technologies being worked on by the Hungarian authors. This research thus contributes to the creation and improvement of fundamental language modules.

Considerable research has been done on automatic identification of collocations and multi-word lexemes in English (Kilgarrif and Tugwell, 2001), German (Kermes and Heid, 2003) and Dutch (Villada & Bouma 2002; Villada, 2004) and so on., Although corpus development has made significant progress in the last few years (Váradi, 2002), no research has been carried out at this level of complexity for Hungarian. Concerning the extraction of collocations, dictionaries have been mined by MorphoLogic's research group for various purposes, including the establishment of the Hungarian nominal WordNet (Prószéky & Miháltz, 2002a, 2002b). In automatic identification of multi-word lexemes for languages lacking a large enough treebank, the most effort is put into pre-processing, that is, preparing the extraction corpora, adding morpho-syntactic annotation and extracting candidate expressions. After pre-processing, thousands of candidate expressions in the datasets are ranked by statistical measures.

2. The resource preparation framework

If detailed linguistic annotation is considered essential for a corpus to be used for collocation statistics, the range of corpora we can use for research is restricted to a great extent. However, if we have access to a tool set capable of extracting collocation candidates from unannotated or sparsely annotated corpora, we can obtain a sufficiently large base corpus in a reasonable time.

We think that collocation research is especially valuable if it aims at finding *typed collocations*, that is, collocations selected on the basis of some morpho-syntactic properties. Examples are the extraction of <verb+PP> pairs, <adjective + noun> bigrams, or <verb + noun + morphological category> trigrams, as opposed to the extraction of *typeless collocations*, ngrams of surface lexemes, like in some contemporary term extraction systems (Castellví et al., 2001).

2.1. Requirements and architecture of the tool set

The tool set described below is designed to extract typed collocation candidates from unannotated corpora, and to prepare data sets for evaluation by the *ngram* statistical package (Pedersen and Banerjee, 2003).[1]

Presently, the configuration of the tool set allows for extracting Hungarian and English collocations only. However, the design of the framework focuses on re-usability and language independence, thus the character sets, the lexicons, and the grammar can be replaced by those of a different language.The tool set implements a sequence of text manipulation and annotation operations. This sequence is split into the following principal phases:

(1) Unified formatting of corpus texts;

(2) Partial (shallow) parsing of texts, and extraction of specific instances, namely, typed collocation candidates;

(3) Heuristic post-filtering of the instances;

(4) Counting the instances and preparation of dataset in the format required by the statistical package.

2.2. Unified formatting of corpus texts

In order to provide a consistent text format, the tool set automatically applies an XML structure that easily accommodates texts at different levels of annotation, let them be unannotated, partially or fully annotated. The XML format retains the most important formatting properties only, such as heading structure, formatting information and heading levels but skipping tables and figures. This pro-

[1] Earlier called NSP, *ngram* is a SourceForge project available at http://sourceforge.net/projects/ngram.

prietary format has been developed during the compilation of the corpus mentioned in Section 3.1. (Kis & Kis, 2003).

2.3. Parsing and Extraction of Candidates

In the phase of extracting collocation candidates, the tool set employs as much natural language processing as possible. The extraction process is capable of deriving bigrams and trigrams from corpus text, all filtered on the basis of morpho-syntactic properties of each unigram. To this end, the tool set incorporates one of two parsers: HumorESK[2] or Moose, both developed by the Hungarian authors, the former mainly for Hungarian, the latter for English texts.

Instead of using specialized scripts, candidates are extracted using a set of metarules. The extraction process employs a post-processor on top of the parser, and those metarules perform specific queries over the parse forests. One can extract relative roots (in principle, any node) of subtrees within the parse forests, and has access to all properties (features) of each node. Thus bigrams and trigrams are composed of parse forest nodes.

Below is an example for a metarule:

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

This metarule – applied in the experiment described in Section 3 – extracts a trigram consisting of the verb's lemma, the lemma of the head of an NP, and the casemark of (the head of) the same NP. Furthermore, the verb and the NP occur within a window of five terminal symbols.

This represents a close co-occurrence, being far smaller than the one used by Villada (ms). The window was deliberately chosen to limit noise because significant noise is already introduced by the lack of disambiguated POS tagging. However, the small window is controversial as the extractor will disregard complex NPs and their collocations.

It is worth noting that, in addition to the surface and lexical forms, the program is capable of extracting any morpho-syntactic feature of a node in the parse tree, such as case endings. This fact becomes crucial in our experiment, where a typical trigram extracted is:

`küld üzenet ACC` *üzenetet küld* 'send a message'

2.4. Heuristic post-filtering of the instances

As already mentioned, a suitable disambiguating POS tagger was still being developed for Hungarian at the time of writing. Thus, the proposed tool set was designed to operate either with or without one.

Without a disambiguating POS tagger, morphological analysis and parsing runs in a single process. Although parsing can be rather deep and the nature of both the parser and the grammar allow for rules (patterns) overriding other rules (to constrain the effects of ambiguous morphological analysis), parsing errors can still occur due to morphological misclassification.

Based on the observation of the error types, we reviewed a number of morphological misclassifications, and developed a filtering mechanism which discards some of the collocation candidates. The 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 number used as a noun and not as a quantifier) 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 precise enough POS tagger is present.

2.5. Counting and preparation of the data set for statistical ranking

The *ngram* statistics package applies statistical functions to frequency data resulting from an independent counting process. Resource preparation must therefore include a process to count instances. We count both ngram (collocation candidate), as well as their components, because the statistical algorithm requires frequencies of the ngrams and each of its composite units.

The tool set presented here includes a robust frequency counting module that was designed for scalability, in order to operate on event sets of millions or billions of instances. During our experiments, we performed a similar investigation on English texts of the British National Corpus, with a text base of one hundred million words, ca. one sixth of the BNC. Frequency counting on the dataset extracted from this subcorpus took ca. 40 seconds on an average PC, under Windows. Neither the corpus nor the dataset were partitioned.

Statistical measures implemented within the *ngram* package require not only the frequency of each candidate ngram, but also the frequencies of the components making up each ngram; in our case, the frequency of each candidate trigram is collected on a par with its unigram counts as well as the counts of all possible bigrams. The frequency counter module can easily perform that task.

Furthermore, the tool set must also include another program that merges the separate frequency lists.

The programs that count frequencies and merge frequency lists are built upon the basis of a linguistic indexer module named 'GammaTrie', developed by Mátyás Naszódi at MorphoLogic (still unpublished): it is the fastest linguistic indexer available to the authors at the time of writing.

3. Extracting Hungarian Multi-word Lexemes: Methods and Evaluation

3.1. Experiment goals

As a working hypothesis, we assumed that parts – composite unigrams – of a multi-word lexeme combine with a better-than-chance frequency, i.e. it is more probable to find them together than we would expect based on their individual frequencies. As mentioned, we investigated <*verb* + *noun* + *casemark*> patterns as candidates, where the noun is the head of an NP, with the casemarker attached as a suffix. This structure has been selected in order to provide results comparable to those acquired by Villada (ms) on Dutch V+PP collocations.

Instead of prepositions, Hungarian adds case endings to the NP's head, which is almost always at the end of the

[2] The name stands for *High-Speed Unification-based Morphology Enriched by Syntactic Knowledge* (Prószéky, 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 (recursive transition networks), and thus its operation was very similar to that of HuMor, MorphoLogic's morphological analyzer.

NP itself; or, uses postpositions directly following the NP head. (Although postpositions are less frequent in Hungarian than casemarkers, we will take them into account in future research.) Dutch and English PPs most often translate into a Hungarian NP with casemarker:

az utca vég-é-n
the street end-POSSESSED-SUPERESSIVE
′at the end of the street'

3.2. The corpus

We had access to the recently compiled, but unannotated SZAK Corpus (Kis and Kis, 2003), a parallel corpus of technical texts, containing ca. 1.2 million words per language, though we used the Hungarian component only. We think that a corpus this small is suitable for testing a statistical procedure in a limited time frame. In addition, this size is large sufficiently large for a technical corpus. During resource preparation, the corpus has been automatically annotated but not disambiguated.

3.3. Statistical measures applied

When selecting statistical functions to apply to our data sets, we relied on the evaluation of the Dutch experiments (Villada, ms[3]). We disregarded those functions that the Dutch evaluation found less accurate, and 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 compare the components of an ngram occurring together against each component occurring independently.

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

The salience measure is an adjustment to the mutual information test. Mutual information 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.

We used the bigram-based version of the log likelihood measure in the *ngram* statistical package. The salience function was also applied on partial bigrams, then the rankings were combined.

3.4. Evaluation

We evaluated the results of ranking the <verb+noun+casemark> candidates in detail, both by the log likelihood and the salience functions. We decided to manually check those candidates ranked among the top 100 either by the log likelihood or the salience functions because we had no hand-tagged data to compare the results to; recall that this was the very first experiment with this procedure and these texts.

Manual checking was carried out by three native speaker judges, voting on each candidate, assigning them an integer score between 1 (worse) and 5 (best). Results were based on measuring the agreement between the human judges. We were then able to make some interesting observations. The two most important ones are outlined below.

3.4.1. Valid or apparently valid multi-word lexemes

Most top ranked candidates are valid collocations: 82 by the salience measure, 76 by the log likelihood ranking, among the top 100.

Most instances of noise among the top ranking candidates can be traced back either to parsing errors or morphological misclassifications. We believe that these errors may be largely eliminated by 'simply' improving the morphological analyzer and the parser itself.

A majority of those could be classified as 'transparent' (around score 4) – 57 by salience, 58 by log likelihood. However, many of them form a *terminological collocation* relating to the technical field of the texts in the corpus (e.g.: "click on the radio button", "open the file"). Most of these arguably transparent collocations are indeed important from the aspect of translation, since these should be translated consistently. Still, a significant number of collocations were classified as 'real multi-word lexemes' whose meaning is entirely non-compositional (See Table 1).

	Topic-independent multi-word lexemes
Salience	25
Log likelihood	18

Table 1. Number of topic-independent multi-word lexemes among the first 100 ranks.

From these results, both statistical methods seem useful for Hungarian. Given our experimental settings, salience seems more reliable as it produces less noise in the 100-best list.

3.4.2 Casemarked NPs as verbal affixes

This second observation provides the strongest evidence in favour of the statistical methods.

In Hungarian, some casemarked NPs act as verbal affixes. Verbal affixes are attached to the verb if and only if they are immediate left neighbours of the verb. When the verbal affix is put by syntax after the verb or when they are not immediate neighbours, the affix appears as a separate word. Furthermore, some morphemes behave in syntax and orthography like verbal affixes, but they look morphologically like nouns. In fact, they are still in the process of being reanalyzed from being the NP in a verbal phrase into being a verbal affix. When written into separate words, they look as a verb followed by an NP with a casemark.

When the latter structures occur as separate words, they are linked very strongly, just as collocations. These are obviously the strongest candidates for multi-word lexemes (see Table 2). However, they are already included in almost all Hungarian dictionaries as single verbs.

Dataset extraction spotted these occurrences as 'verb, noun phrase, casemark' collocations, and statistics ranked these instances very high. The ranks in the table might be misleading: it is even more convincing to point out that these collocations are in fact the first occurrences of each verb in the ranked dataset, according to both measures.

[3] Villada, Begona (ms). Acquisition of Dutch support verb collocations: a model comparison. Ongoing work, see http://www.let.rug.nl/~begona/papers/svcmodels.ps.

	Salience	Log-likelihood
hoz lét SUB (*létrehoz*) ‘create’	1	1
vesz ész SUB (*észrevesz*) ‘notice’	8	12
jˏn lét SUB (*létrej ˏ n*) ‘come into being’	13	17

Table 2. Affixed verbs written as separate words

4. Suggestions for Improvement

In order to improve the reliability of the statistical extraction process, we are dealing with three issues at the time of writing:

(1) The accuracy of the morphological analysis of the corpus text must be increased both by introducing a disambiguating POS tagger, and improving the Hungarian morphological analyzer in general.

(2) Larger and less specialised corpora are being prepared for use in addition to the current technical corpus.

(3) The Hungarian parser should be used more efficiently. The present extraction scheme uses the 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 candidate, are in fact children of the same VP node. We suppose that some of the ‘false argument’ errors can be eliminated this way.

Dataset extraction is also being improved through detailed corrections: the proper treatment of verbal affixes (being presently discarded by the parser at a lower level) and the enlargement of the case categories with nominal postpositions

4.1 Introducing further methods

There are fields of collocation research where both rule-based and statistical methods are applied. Real strength lies with the combination of the two. As we have access to many bilingual dictionaries and are working to develop various translation tools, it is obvious 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.

We would require a parallel corpus aligned at the sentence level as a minimum, and a high-quality bilingual dictionary. Taking an ngram, such as an NP or a <Verb, NP, casemark> pattern, from the text in the language under investigation, if there is at least one word in it whose obvious (dictionary-based) translation cannot be found in the alignment pair, such ngram will be a good candidate for further testing.

5. Conclusion

We have first introduced a robust, flexible and reusable resource preparation tool set for the purpose of corpus-based collocation research. From June 2004, a demonstration version of the tool set is downloadable from the MorphoLogic website (www.morphologic.hu/research).

Subsequently, we presented an experiment on the automatic identification of Hungarian multi-word lexemes, demonstrating that both the resource preparation tool set and the statistical measures are suitable for the task, and serve as a powerful starting point for corpus-based investigation of Hungarian collocations.

6. 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

Castellví, M. Teresa Cabré & Bagot, Rosa Estopà & Palatresi, Jordi Vivaldi (2001). Automatic Term Detection: A Review of Current Systems. In Bourigault, Didier & Jacquemin, Christian & L’Homme, Marie-Claude (Eds.), Recent Advances in Computational Terminology (pp. 53–88.) John Benjamins, Amsterdam-Philadelphia.

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

Jacquemin, Christian (2001). Spotting and Discovering Terms through Natural Language Processing. The MIT Press.

Kermes, H. & U. Heid (2003). Using chunked corpora for the acquisition of collocations and idiomatic expressions. In Proceedings of COMPLEX 2003. Budapest.

Kis, Ádám & Kis, Balázs (2003). A Prescriptive Corpus-based Technical Dictionary. Development of a multi-purpose technical dictionary. In Proceedings of COMPLEX 2003, Budapest.

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

Pedersen, T. & S.Banerjee (2003). The Design, Implementation and Use of the Ngram Statistics Package. In Proceedings of the 4th International Conference on Intelligent Text Processing and Computational Linguistics. Mexico City.

Prószéky, Gábor (1996): Syntax As Meta-morphology. Proceedings of COLING-96, Vol.2, 1123-1126. Copenhagen, Denmark.

Prószéky, Gábor (1999). Lexical Information and Decisions in Parsing. In Cristea, Dan, Dan Tufis, Amalia Todirascu, Valentin Tablan & Catalina Barbu (Eds.) 4th Eurolan Summer School on Human Language Technology, Technical Report 99-02, Iasi, Romania.

Prószéky, G., M. Miháltz (2002a): Automatism and User Interaction: Building a Hungarian WordNet. Proceedings of the Third International Conference on Language Resources and Evaluation, 957-961. Las Palmas, Spain.

Prószéky, G., M. Miháltz (2002b): Semi-automatic Development of the Hungarian WordNet. Proceedings of the Workshop on WordNet Structures and Standardization and How These Affect WordNet Applications and Evalutation, 42-46. Las Palmas, Spain.

Váradi, Tamás (2002): The Hungarian National Corpus. Proceedings of the Third International Conference on Language Resources and Evaluation, 385-389. Las Palmas, Spain.

Villada, Begona & Bouma, Gosse (2002): A corpus-based approach to the acquisition of collocational prepositional phrases. In Proceedings of EURALEX 2002, Copenhagen, Denmark.

Villada, B. (2004): Discarding noise in an automatically acquired lexicon of support verb constructions. In Proceedings of LREC 2004, Lisbon. (To appear).