Verb-Noun Collocation SyntLex Dictionary – Corpus-Based Approach

GrażynaVetulani[1], Zygmunt Vetulani[2], Tomasz Obrębski[2]
Adam Mickiewicz University, Poznań, Poland
[1]Faculty of Neophilology
Al Niepodległości 4
[2]Faculty of Mathematics and Computer Science
ul. Umultowska 87
61-614 Poznań
gravet@amu.edu.pl,
vetulani@amu.edu.pl
obrebski@amu.edu.pl

Abstract

The project presented here is a part of a long term research program aiming at a full lexicon grammar for Polish (SyntLex). The main concern of this project is computer-assisted acquisition and morpho-syntactic description of verb-noun collocations in Polish. We present methodology and resources obtained in three main project phases which are: dictionary-based acquisition of collocation lexicon, feasibility study for corpus-based lexicon enlargement phase, corpus-based lexicon enlargement and collocation description. In this paper we focus on the results of the third phase. The presented here corpus-based approach permitted us to triple the size the verb-noun collocation dictionary for Polish. In the paper we describe the SyntLex Dictionary of Collocations and announce some future research intended to be a separate project continuation.

1. Introduction

The project presented here is a part of a long term research program aiming at development of a full lexicon-grammar for Polish (SyntLex). By lexicon-grammar we mean, after Maurice Gross (1975), the theory of syntax based on elementary sentences of natural language. The major principle is that the meaning of the word (predicative noun or verb) is defined by elementary sentences in which this word occupy the predicative position. It follows that dictionary descriptions of a (predicative) word should (explicitly or implicitly) contain generative information about the kind of sentences that may be formed around this word. Dictionaries respecting this principle have proved their utility for designing natural language parsers. For example, organization of grammatical information in the form of lexicon-grammar permitted us to implement an efficient heuristic parser of Polish sentences being a part of the POLINT system (Vetulani 1997).

This paper focuses at acquisition and morpho-syntactic description of verb-noun collocations in Polish. We present methodology and resources obtained in three main project phases. In particular, we will describe the SyntLex Dictionary of Collocations and announce some future research (a separate on-going project).

The three main phases of the project were:

1) Dictionary-based acquisition of collocation lexicon (called Basic Resource, BR, cf. below)
2) Feasibility study for corpus-based enlargement of the BR
3) Corpus-based lexicon enlargement and collocation description

The first two phases were already presented in (Vetulani et al. 2006, 2007), therefore, we will limit ourselves to short overview of the main aspects.

2. The first phase: paper dictionary based acquisition of collocations

The dictionary-based phase consisted in extracting collocations from the existing, traditional dictionaries by a linguist-lexicographer (Vetulani, G. 2000). For this purpose, the lexicographer inspected available dictionaries of Polish. This part of the work was based first of all on "Dictionary of Polish" (Szymczak, 1978) which is the main reference dictionary with over 100,000 entries. The inspection of 40,000 nouns permitted to to classify 7500 predicative nouns. Five subclasses have been identified, the first of them amounting to 2862 lexemes representing names for various kinds of activities and behavior (e.g. names of actions, procedures, ...) contains nouns forming collocations with a large number of support verbs in a quite irregular way.

The remaining four classes (abstract properties, names of diseases, names of professions, nouns supported by so called occurrence verbs) are much more regular and their facility to form collocations with various kinds of verbs are much more limited. These predicate nouns often share the support verbs with a large number of other class members. We therefore focused our attention on the first class, the most collocation-productive.

The set of entries corresponding for this class is called Basic Resource (BR). An entry of the BR presents collocations which may be formed for the given predicative noun using different support verbs (simple or compound). The annotation provides substantial morpho-syntactic information constituting the valency scheme of the collocation. It provides also formal properties of components (case, obligatory prepositions if any). Here follows an example of an entry (a fragment of).

Example 1.
rozmowa, f /
nawiązać (Acc)/N1 z (Instr),
odbyć(Acc) / N1 z (Instr), ...[1]

As one may see in the Example 1, as well in the Example 2 below, one entry typically contain more then one collocation and morpho-syntactic features may be different depending on the support verb. Also, meaning may change depending on the support verb[2]. In the Example 2 *przeprowadzić amputację, dokonać ampuracji* both mean to perform amputation, whereas *poddać się amputacji, ulec amputacji* means *be amputed (of sth).*

Example 2. BR items.
...
aluzja, f/
allusion
czynić(Acc,pl)/N1do(Gen),
to make an ~ to sth
robić(Acc)/N1do (Gen),
to make ~s to sth
pozwalać sobie na(Acc,pl)/N1do(Gen)
to dare to make ~s to sth

ambaras, m/
embarrasment
mieć(Acc)/N1z(Instr)
to have an ~ with sb

ambicja, f/
ambition
mieć(Acc)/N1(Gen),
to have an ~ of sth
mieć(Acc,pl)/MOD
to have MOD ~s

amnestia, f/
amnesty
ogłosić(Acc),
to declare ~
uchwalić(Acc),
to vote ~
ustanowić(Acc)
to declare ~

amok, m/
amok
być w(Loc),
to be in ~
dostać(Gen),
to get ~
wpaść w(Acc)
to fall into ~

amory, m;pl/
amours
wdać się w(Acc)/N1z(Instr)
to enter into ~ with sb

amputacja, f/
amputation
przeprowadzić(Acc)/N1(Gen),
to carry an ~ of sth
dokonać(Gen) /N1(Gen),
to perform an ~ of sth
poddać(Dat)/N1(Acc),
to subject sth to an ~
poddać się(Dat)/N1 (Gen),
to undergo an ~ of sth
ulec(Dat)
to undergo an ~ of sth

anabioza, f/
anabiosis
poddać się(Dat),
to undergo an ~
ulec(Dat)
to undergo an ~
...

In the examples above, the symbol MOD indicates the obligatory occurrence of an adjectival modifier of the predicate noun.

3. The second phase: feasibility study

The second phase consisted in:

- proposition of an algorithm for computer-assisted corpus based acquisition of new collocations (Vetulani et al., 2006), where by "new" collocations me mean those attested in a corpus, but absent in the BR,
- feasibility study (Vetulani et al., 2007).

One of important drawbacks of the BR consisted in lack of many collocations of practical importance. This was due to the absence in traditional dictionaries of many commonly used collocations (typical for human addressed dictionaries). Therefore we proposed computer-assisted enlargement method (Vetulani et al. 2006, 2007) in the form of a processing algorithm permitting a substantial reduction of manual search in the corpus. We also effected 5% feasibility study using the IPI PAN Corpus (Przepiórkowski, 2004). This study consisted in application of the algorithms to a random selected 5% sample of the BR list of predicate nouns. We concluded (Vetulani et al. 2007) that procedures when applied to the whole BR would most probably result with a substantial number of "discoveries" (collocations found in the corpus but absent in the BR) for a very reasonable amount of effort evaluated to 5-10 man-month[3]. On the basis of the experiment we estimated

[1] rozmowa = conversation, nawiązać rozmowę = enter into conversation, z = with, odbyć = to take place, odbyć rozmowę = to have conversation; f-feminine, Acc-accusative, Instr-instrumental, N1 = argument position opened by the predicate noun, 'N1 z (Instr)' = the argument at the position N1 is composed of the preposition 'z' and a nominal group in instrumental case

[2] Which means that in many cases support verbs are not semantically void (this issue has been discussed by Vetulani, G. (2000)).

[3] One of the main cost-generating factors is the time necessary to read the corpus, especially if we were constrained to inspect it

that the final set of collocations would be at least twice as large as the initial one[4].

The following example shows the result obtained for the predicative noun "ambicja" (*ambition*).

Example 3. Predicate noun "ambicja" (*ambition*):

A. The entry "ambicja" in the Basic Resource
ambicja, f/
ambition
mieć(Acc)/N1(Gen),
to have an ~ of sth
mieć(Acc,pl)/MOD
to have MOD ~s

B. The entry "ambicja" including collocations retrieved in the corpus (in italics)

ambicja, f/
ambition
mieć(Acc)/N1(D),
to have an ~ of sth
mieć(Acc,pl)/MOD,
to have MOD ~s
posiadać(Acc,pl)/MOD,
to own MOD ~s
ujawniać(Acc,pl)/MOD,
to show MOD ~s
zaspokoić(Acc)/N1(Gen),
to fulfill one's ~ of sth
zaspokoić(Acc,pl)/MOD,
to fulfill MOD ~s
zaspakajać(Acc)/N1(Gen)
to fulfill one's ~ of sth
zaspakajać(Acc,pl)/MOD
to fulfill MOD ~s

It is interesting to see that among "discoveries" it is possible to find very common collocations like *posiadać* (*to own*) *ambicje*, or *ujawniać (to show) ambicje.*

4. The third phase: corpus based dictionary enlargement

The solution we proposed for the BR enlargement phase consists in semi-automatic (machine-assisted) transformation of the corpus in order to reduce human processing effort (time). The resources (data and tools) we used were:

- the Basic Resource in electronic form
- a large size corpus (a part of the IPI PAN Corpus) (Przepiórkowski 2004)
- tools in the form of processing software (lemmatizer, concordancer,...) (We used our own toolkit described in (Obrębski&Stolarski 2006)).

The processing was structured into a sequence of five consecutive steps.

Step 1. (automatic)
Application of the corpus filtering procedures to the input data (corpus, basic resource). (Cf. (Vetulani et al. 2006, 2007) for the detailed description of the corpus filtering procedure.) The objective of this step was to extract corpus fragments likely to contain occurrences of predicative constructions. From this data a list of support verb candidates for each noun was created.

Step 2. (manual)
The lists of support verb candidates for each noun were analysed by lexicographers. The main objective of this task was to shorten the lists obtained in Step 1 by eliminating the verbs for which the negative decision could be made without context inspection.

Step 3. (automatic)
For all noun-verb pairs retained in the output of Step 2 the whole expressions matching the general pattern of predicative constructions, together with fragments of their left and right contexts were extracted from the corpus in the form of concordance tables.

Step 4. (manual)
The noun-verb pairs were processed manually by lexicographers. Their task consisted in:

a) checking whether the corpus samples confirm the predicative use of the noun-verb pair,
b) providing morpho-syntactic descriptions of the noun-verb pairs observed in the corpus (schemata of the predicative constructions in the format used in BR (cf. Example 1 and 2, above)),
c) choosing a corpus sample(s) best illustrating the use of the predicative construction(s) based on the noun-verb pair.

Step 5. (manual)
Verification and correction by senior experts or/and lexicographers and final formatting. tables.

Steps 1-5 constitute, basically, application of the algorithm described in (Veulani et al., 2006, 2007). The most important modification was the elimination of the intermediate step consisting in examining noun-verb pairs first in a very restricted context in order to reduce the size of the data processed in Step 4. We realized that the gain due to the data size reduction would not compensate the introduction of additional tasks (the overhead related to problems of organizational and logistic nature were not identified while working with a small sample).

manually. The trade-off between cost and benefice makes a part of the game in empirical, corpus-based linguistics (language engineering). For the purpose of the project we were allowed to use a part of the IPI PAN Corpus containing env. 80,000,000 of text words. This corpus size correspond to ca 100,000 printed pages (considering 800 words per page). With the speed of 10 pages processed manually per day, the reading of the total corpus would require 10,000 working days, i.e. ca 500 man-months. This would correspond to the involvement of a 10-people full-time team in a 4 year project based on mainly manual processing (reading of the corpus).

[4] In a fact, the gain was much higher, cf. later.

5. Conclusions

The implementation of the algorithm for the total of BR appeared successful. To support this claim we propose to consider the following: size/quality increase, costs, positive side effects.

a. Size/quality increase

The number of 5404 collocations of the BR (i.e. for 2862 entries) grew up to ca 16,000. The new collocations retrieved from the corpus may reasonably be qualified as "important".

b. Effort

The cost of the project enlargement measured in effort appeared to confirm the evaluation made on the basis of the 5% feasibility test. This cost amounted to ca 8 man-months (6 man-month - extraction, 2 man-month - verification and final formatting of the results).

c. Additional effects

An important side-effect of the applied method was the possibility to provide usage examples extracted from the corpus, as well as to make interesting quantitative observations (which are being currently the object of separate studies). Limitations[5] of the applied method have also been observed: 675 (out of 2862) of predicative nouns from the BR list were not found in the IPI PAN corpus at all or at least in the context matching the predicative construction pattern. This automatically rises the question of the quality of the applied corpus[6]. For the remaining 2187 nouns over 1700 collocations listed in BR were not attested in the corpus.

6. Future research

The completion of this BR enlargement project permitted us to start the next project aiming at integration of the SyntLex collocations into the PolNet.[7]. We intend to use the SentLex as well as the PolNet based ontology in the Polish language understanding module of the POLINT-112-SMS system (Vetulani 2007), an application to communicate in emergency situations via text messages.

[5] The completeness of linguistic resources collected for further processing or application design represents always a delicate problem. This is so because language systems are known to be (usually) open, potentially infinite and productive. Therefore it is practically impossible to imagine any exhaustive listing of occurrences for most of interesting language phenomena. Corpus-based methodologies try to cope with these problems but are costly and have limited coverage (as not all phenomena may appear in corpora for various reasons).

[6] It is to be noticed, however, that we have used the best corpus available at the time of the project.

[7] PolNet is an on-going wordnet project for Polish, being realized within Polish Platform for Homeland Security at the Adam Mickiewicz University (Vetulani et al. 2007a) since 2007. The results presented in this paper will directly help extending the PolNet to cover also verbal synsets.

7. Acknowledgements

This work has been supported by the Polish Ministry of Science and Higher Education (Ministerstwo Nauki i Szkolnictwa Wyższego) grant R00 028 02 for the period 2006-2009 (within the Polish Platform for Homeland Security).

We wish to thank the Institute of Computer Science of the Polish Academy of Sciences, Warsaw, for kindly providing us with the IPI PAN Corpus necessary for this research.

8. References

Gross, M. (1975). *Méthodes en syntaxe,* Paris: Hermann.

Obrębski, T., Stolarski, M. (2006). UAM Text Tools - a flexible NLP architecture. In: N. Calzolari (ed.), *Proceedings of the Fifth International Conference on Language Resources and Evaluation, Genoa, Italy, 24-26.05.2000*, ELRA, Paris, pp. 2259--2262

Przepiórkowski, A. (2004). *The IPI PAN Corpus*, IPIPAN, Warszawa.

Szymczak, M. (ed.) (1978). *Słownik języka polskiego.* (Dictionary of Polish Language; in Polish).

Vetulani, Z. 1997. A system for Computer Understanding of Texts, in: R.Murawski, J. Pogonowski (eds), *Euphony and Logos (Poznań Studies in the Philosophy of the Sciences and the Humanities, vol. 57*), Rodopi, Amsterdam-Atlanta, 387--416.

Vetulani, G. (2000). *Rzeczowniki predykatywne języka polskiego* (Predicate Nouns of Polish; in Polish), Wyd. Nauk. UAM, Poznań.

Vetulani, Z. (2007). Natural language based communication between human users and emergency center in critical situations. POLINT-112. In Z. Vetulani (ed*.) Proceedings of the 3rd Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, October 5-7, 2007, Poznań, Poland,* Wyd. Poznańskie, Poznań, pp. 571--572.

Vetulani, Z., Obrębski, T., Vetulani, G. (2006). Syntactic Lexicon of Polish Predicative Nouns, in: N. Calzolari (ed.), *Proceedings of the Fifth International Conference on Language Resources and Evaluation, Genoa, Italy, 24-26.05.2006*, ELRA, Paris, 1734-1737.

Vetulani, Z., Obrębski, T., Vetulani G. (2007). Towards a Lexicon-Grammar of Polish: Extraction of Verbo-Nominal Collocations from Corpora. In: *Proceedings of the Twentieth International Florida Artificial Intelligence Research Society Conference (FLAIRS-07)*, AAAI Press, Menlo Park, California, pp. 267--268.

Vetulani, Z., Walkowska, J., Obrębski, T., Konieczka, P., Rzepecki, P., Marciniak, J. (2007a). PolNet - Polish WordNet project algorithm. In: Z. Vetulni (ed.) *Proceedings of the 3rd Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, October 5-7, 2007, Poznań, Poland*, Wyd. Poznańskie, Poznań, pp. 172--176.