

# Towards a Lexicon-Grammar of Polish: Extraction of Verbo-Nominal Collocations from Corpora

Zygmunt Vetulani, Tomasz Obrębski , Grażyna Vetulani

Adam Mickiewicz University  
ul. Wieniawskiego 1, 61-712 Poznań, Poland  
{vetulani,obrebski.gravet}@amu.edu.pl

## Abstract

In the paper we present a contribution to the SyntLex long-term-project aiming at a lexicon-grammar for Polish. A corpus-based method is presented for computer-assisted improvement or/and verification of verbo-nominal lexicon-grammars (in application to Polish). Feasibility study.

## Introduction

The lexicon-grammar-based paradigm attributes the central role to the predicative constructions. The collection of verbo-nominal collocations with the verb playing the role of syntactic support and the noun carrying predicative information constitutes (in Slavonic languages, but also in, e.g., French or English) an important part of the class of predicative expressions.

In this paper we present a part of our SyntLex project aiming at the full lexicon-grammar for Polish. Our starting point was the lexicon of verbo-nominal collocations collected by Vetulani G. (2000). To obtain this lexicon (called Basic Resource, BR), traditional dictionary-based (manual) methods were used. The well known drawbacks of the traditional dictionaries are:

- for many predicative nouns (i.e. lexicon entries) some of the frequently used collocations were not considered,
- the selection of lexical entries (list of the words) does not reflect the usage of the words,
- important omissions and gaps are due to some extra-linguistic factors, e.g. omission of vulgar terms, slang words, size limitation ...

We decided to complete the systemic gaps that are related to the nature of traditional dictionaries by applying corpora-based collocation acquisition techniques permitting direct access to the linguistic phenomena manifested in texts. For this task we use a non-annotated fragment (80 mln words) of the IPI PAN corpus of Polish (Przepiórkowski 2004). We focus on the acquisition of

new collocations for already identified predicative words and on corpus-based confirmation of the collocations already collected by G. Vetulani (2000). The processing scheme in its preliminary form was presented in (Vetulani et al. 2006). Below, we present the final version of proposed computer-assisted processing methods together with the results of “feasibility tests” performed on a BR-extracted sample.

## Basic Resource

The BR contains systematic description of predicative nouns. It was obtained through thorough investigation of ca. 40000 nominal entries in a classical (paper) dictionary. 7500 nouns were classified as predicative and annotated with syntactic information. A subclass of the BR consisting of 2826 abstract names of various kinds of activities and behaviour is of particular interest for lexicon-grammar construction (Vetulani G, 2000). This subclass will be called BR<sub>1</sub>.

### Example 1. BR<sub>1</sub> items

amnestia, f/ ogłosić(Acc), uchwalić(Acc), ustanowić(Acc)  
[amnesty/ announce, vote, declare]

## The processing scheme

It is well known that corpus based methods which require manual work are very time consuming. With the size of today's corpora the reading (hand processing) time is one of the main bottlenecks. Our challenging objective consisted in finding language engineering techniques to make this effort affordable (and to evaluate this effort). Our aim was to discover the corpus-attested verbo-nominal collocations for a given set of predicate nouns.

The first task consisted in developing a corpus filtering method in order to minimize the amount of text destined to be read by humans. The filter was supposed to extract occurrences of (potential) predicative nouns in their context covering the support verb and complements of the

noun (see (Vetulani Z. et al. 2006) for a discussion of alternative pattern forms). The filtering consisted in extraction of concordances for predicative nouns, following the general pattern:

<verb> + <predicate\_noun> + [<noun> + [<noun>]]

Additionally, prepositions before nouns, the negation particle (“nie”) before the support verb, the reflexive pronoun “się” before and after the verb, as well as the adjectival modifiers before predicate noun were admitted.

Naturally, only some of the extracted expressions contain bona fide predicative nouns. In consequence, only a part of verbs accompanying the noun could be classified as playing the role of a support verb.

The distinction of predicative and non-predicative use could not be made automatically. Therefore the output of the filtering procedure had to be scanned manually.

### Feasibility Experiments (5%-sample based)

We will describe here in some detail an experiment aimed at estimating the effort necessary to apply the procedure to the BR as well as estimating the results possible to obtain with the adopted approach. The experiment was conducted on a 5%-sample of BR<sub>1</sub>.

Step 1. Application of the filter patterns described above. The resulting file A had the size of 61250 lines (1072 pages, 235742 words).

Step 2. Deletion of the right contexts of the predicate noun, transformation of verbs into the infinitive form and elimination of repetitions. The resulting file B contains 27393 lines (452 pages, 82 464 words).

Step 3. Deletion of all the elements except the noun and the verb, transformation of the noun into its base form and elimination of repetitions. The resulting file C contains 12628 lines (210 pages, 12 485 words).

Step 4. Manual processing of the file C in order to discover *collocation candidates*.

Step 5. For the retrieved *collocation candidates* the (partial) context has been resumed (from B). The resulting file B' contains 4324 lines (72 pages, 12 783 words).

### Example 2. Samples of files A, B, C

```
A:
ambicja
 16      ma [ambicje]
  2      zaspokoi [ambicje] kolejnych koalicji

B:
ambicja
 35      <mieć> [ambicje]
  8      <zaspokoić> [ambicje]

C:
ambicja
 79 mieć
  9 zaspokoić
```

## Experiment Results

**Effort estimation.** On the basis of the experiment we may evaluate the total reading effort at ca 4 man-month under the assumption that each text is being read by one lexicographer. Considering that human expertise (language intuition) is involved, application of two (or more for some steps) human experts is advisable. The total cost of the project (without considering elaboration of processing tools) should be estimated at 5-10 man-months.

**Resource extension.** The experiment has shown practical usefulness of the methods resulting in interesting discoveries of collocations not attested in the BR. On the basis of the experiment we estimate that the final set of collocations will be twice as large as the initial one (BR). The following example shows the result of the procedures described in the paper as applied to the predicative noun “ambicja” (*ambition*).

### Example 3. BR entries

(i). The entry “ambicja” in the Basic Resource

ambicja, f/ mieć(Acc)/N1(Gen), mieć(Acc,pl)/MOD

(ii). The entry “ambicja” including collocations retrieved in the corpus (in italics)

ambicja, f/ mieć(Acc)/N1(D), mieć(Acc,pl)/MOD,  
*posiadać*(Acc,pl)/MOD, *ujawniać*(Acc,pl)/MOD,  
*zaspokoić*(Acc)/N1(Gen), *zaspokoić*(Acc,pl)/MOD,  
*zaspakajać*(Acc)/N1(Gen),  
*zaspakajać*(Acc,pl)/MOD

Notice that the method described above as a way to improve a resource obtained within different methodological paradigm (dictionary-based) may be used to develop collocation dictionary from scratch (for the initially pre-selected set of predicate nouns).

## Acknowledgements

This research is partially covered by Polish Government research grant R00 028 02 for the period 2006-2009.

## References

- Przeźiórkowski, A. 2004. *The IPI PAN Corpus*, IPIPAN, Warszawa.
- Vetulani, G. 2000. *Rzeczniki predykatywne języka polskiego* (Predicate Nouns of Polish; in Polish), Wyd. Nauk. UAM, Poznań.
- Vetulani, Z., Obrebski, T. and Vetulani, G. 2006. Syntactic Lexicon of Polish Predicative Nouns. In: N. Calzolari (ed.), *Fifth International Conference on Language Resources and Evaluation*, Genoa, Italy, 24-26.05.2006, (Proceedings), ELRA, Paris, 1734-1737.