Syntactic Lexicon of Polish Predicative Nouns

GrażynaVetulani[1], Zygmunt Vetulani[2], Tomasz Obrębski[2]

Adam Mickiewicz University
[1]Faculty of Neophilology
al. Niepodległości 4
60-665 Poznań, Poland
[2]Faculty of Mathematics and Computer Science
ul. Umultowska 87
61-614 Poznań, Poland
{grazyna.vetulani,zygmunt.vetulani,tomasz.obrebski}@amu.edu.pl

Abstract

In the paper we report realization of SyntLex project aiming at construction of a full lexicon grammar for Polish. The lexicon-grammar based paradigm in computer linguistics is derived from the predicate logic and attributes a central role to the predicative constructions. An important class of syntactic constructions in many languages (French, English, Polish and other Slavonic languages in particular) are those based on verbo-nominal collocations, with the verb playing a support role with respect to the noun considered as carrying the predicative information. In this paper we refer to the former research by one of the authors aiming at full description of verbo-nominal predicative constructions for Polish in the form of an electronic resource for LI applications. We describe procedures to complete and corpus validate the resource obtained so far.

1. Introduction

The study reported aims at corpus-based validation and improvement of the so called *Basic Resource of Polish predicative nouns forming verb-noun collocations (*shortly *Basic Resource)* built within the project of Syntactic Lexicon of Polish Predicative Nouns. This is an important part of the long-term research program aiming at the Lexicon Grammar for Polish. In this paper (Chapter 2) we present the *Basic Resource* (being now prepared for presentation and public distribution for research purposes through the ELRA/ELDA organisation). In the rest of the paper we will describe the method implemented in order to improve the essential part of the *Basic Resource*. Namely, in Chapter 3 we will comment on the shortcomings of the Basic Resource and in the crucial Chapter 4 we will present language engineering methods applied in order to make an improvement of the BR feasible.

2. Basic Resource

The *Basic Resource* is a result of systematic analysis of a large set of Polish predicative nouns and their usage. More than 40,000 nouns were investigated: the possibility of forming nominal predicate (together with some support verb) was studied and possible structures were described (Vetulani, G. 2000). This work brought about a selection of ca 7500 entries clustered into 5 classes of different size and homogeneity. This clustering takes into consideration the relationship between support verbs and predicate nouns. The Class I (2878 entries) contains nouns that are names of for various kinds of activities and types of behaviour (names of usual and unusual actions, procedures, techniques, methods, operations, states, processes, etc.). All these nouns select their owns arguments and the support verb (several predicate nouns may have the same support). There is however no common set of support verbs shared by all members of this class. Support verbs may be simple or compound, neutral or marked (concerning style or aspect). Here follows an example of an entry (a fragment of):

rozmowa, f / nawiązać (Acc), N1 z (Instr), odbyć(Acc) / N1 z (Instr), ...

(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)

Besides Class I discussed above Vetulani (Vetulani, G., 2000) considers four other classes. Classes (II-V) are more homogenous from the point of view of their links with arguments and the support verb. The predicate nouns often share the support verbs with a large number of other class members. Also, the number of typical support verbs is relatively restricted.

Class II contains abstract nouns standing for properties (features) (2826 items). For this and the remaining classes, the support verbs are not explicitly associated in the *basic resource*, although the list of the most typical ones was identified for the Class II. To complete this gap will be one of the main objectives of the further research. The Class III contains names of diseases (275), the Class IV - names of professions (1478) and finally the Class V contains nouns supported by the so called occurrence support verbs (212)

3. Why to improve the Basic Resource?

The *Basic Resource*, despite its formalised character and careful description, has shortcomings resulting from the applied methodology and more precisely from the way the Basic Resource was obtained (as a result of manual processing of traditional dictionaries). E.g. we observe that some frequently used collocations are absent in the Basic Resource and a number of archaic and rare collocations without practical significance are included.

In the fragment of the project presented here, we focus on the Class I, composed of 2878 words. For this class we propose a machine-assisted method of Basic Resource improvement. As we have observed above, the Basic Resource was compiled (Vetulani, G.) on the basis of

dictionary research and some former lexicographical works (Jędrzejko, 1998) and so far was not directly verified in a systematic way with empirical data in the form of a large corpus. The selection of collocations is therefore biased by lexicographical choices of the dictionary authors and other researchers. Availability of a large corpus for Polish with ca 80 000 000 words (Przepiórkowski 2004) made it possible to confront the Basic Resource with this corpus.[1] *A priori*, results of such a confrontation are interesting both for the Basic Resource developers (the authors of this paper) and the corpus designers.

As we shall see, exploring the corpus may considerably help to complete gaps in the Basic Resource. These gaps consist in the frequent absence of collocations in dictionaries. We must however admit that even a large, high quality text corpus may not guarantee the full coverage with respect the dictionary content. Indeed, for the 2878 items (entries) of the Basic Resource Class I, only 2406 were found in the IPI PAN corpus (the largest one for Polish at the moment). Of course this drawback should be considered as a temporary limitation, not affecting the validity of the methodology we propose. (A systematic study of this phenomenon could possibly help us to better understand what the real coverage of the IPI PAN corpus is. We will not pursue this track here and will limit ourselves to the mere observation that the coverage gap concerns mainly a very sophisticated part of the vocabulary, in most cases of scientific character.) What we expect to obtain from the systematic "mining" into the corpus is the completion of the predicative nouns descriptions by the discovery of new support verbs for nouns already identified as predicative. It is hard to expect that this objective may be achieved in a fully automatic way because this would have to involve high level processing methods based on very efficacious tools and solid language resources, whereas this research is intended as a contribution *towards* development of such tools. On the other hand, looking for machine-assisted tools and methods is not only a methodological option but a strict necessity as purely hand processing of a large corpus is too time consuming.[2] What we expect from the language engineering is such a transformation of the data (corpus) that would substantially reduce the reading time. In what follows we will describe a set of steps aiming at such an effect.

4. Procedure

In this Chapter we describe the algorithm of machine-assisted corpus processing that we propose to apply in order to improve the Basic Resource. The main idea of this algorithm is to transform the corpus data in such a way to substantially reduce the search time with respect to that of manual corpus processing.

4.1. Input resources

The input to the algorithm consists of the following resources:

1) Basic Resource consisting of 2878 entries for Class I predicative nouns from (Vetulani, G. 2000).

2) The publicly available part of the IPI PAN corpus (Przepiórkowski, 2000) without morphological (and any other kind of) annotations.

4.2. Preparatory steps

The search operations are preceded by preparatory processing of the input resources. These were:

1) Extraction of the list (L) of BR predicative nouns from the the Basic Resource (2878 words),

2) Extraction of the list of BR collocations (K) of the form support_verb+predicative_noun (5123 collocations),

3) Extraction of the list of verbs (SV) already identified as supports in the BR collocations (495 verbs),

4) Extraction of the set of structural schemes (S) for the BR extracted collocations (5404 schemes).

This last step consists in transformation of the Basic Resource items into structural schemes.

Transformation example.

A Basic Resource item:

przyjaźń, f / darzyć (Instr) / N1 (Acc), okazać (Acc) / N1 (Dat) / N1 (Dat), dochować (D) / N1 (D)

Derived schemes:

darzyć + przyjaźń (Instr) + <noun> (Acc)

okazać + przyjaźń (Acc) + <noun> (Dat)

dochować + przyjaźń (Gen) + <noun> (Dat)

The schemes derived correspond to the following collocations (respectively): *darzyć przyjaźnią (to grant friendship), okazać przyjaźń (to manifest friendship), dochować przyjaźni (to remain friend)*. The surface ordering of the element may be different that the ordering of the corresponding scheme (which represents the so called neutral ordering).

5) Transformation of structural schemes into patterns at three abstraction levels.

a) Replacement of the verb at the support position by the variable <verb> (without modifying the predicate noun) will result with the set of structural schemes which may be considered as a direct generalization of the observed collocation for the given predicate noun:

<verb> + przyjaźń (Instr) + <noun> (Acc)

<verb> + przyjaźń (Acc) + <noun> (Dat)

<verb> + przyjaźń (Gen) + <noun> (Dat)

It worth noticing that for each predicate noun only a small number of corresponding schemes was observed.

b) The next generalization consists in replacing the predicate noun by a variable <predicate_noun>. This operation gives a set of patterns (P):

<verb> + <predicate_noun> (Instr) + <noun> (Acc)

<verb> + <predicate_noun> (Acc) + <noun> (Dat)

<verb> + <predicate_noun> (Gen) + <noun> (Dat)

It is remarkable, that the number of such patterns is relatively small with respect to the number of collocations identified in the BR (257/5123).

c) One more generalization step will be useful. These most general schemes will be derived from the above

[1] As a matter of fact, the IPI PAN corpus is substantially larger, but only a part was made publicly accessible.

[2] To better see the problem, let us consider the IPI PAN corpus with ca 80 000 000 of words (i.e. the part of the corpus which is publicly available). This corpus size correspond to ca 100 000 printed pages (considering 800 words per page). With the processing speed estimated to 10 pages 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).

patterns by: making abstraction of inflection cases, admitting prepositions before nouns, negation particle ("nie") at the positions before the support verb, "się" at the positions before and after the verb, as well as the adjectival modifiers before or after predicate noun. For the obvious reasons, the number of such patterns is very limited (56).

The optional solution we considered was to add an additional step (as *the first* exploration step of Chapter 4.3.) consisting in extracting concordances from the corpus in order to provide context surrounding collocations in the corpus. Then some procedures (manual or machine) are to be engaged in order to *discover* new patterns, i.e. patterns which are not derivable from the Basis Resource. Such a pattern discovery step is time consuming, and we claim that it may be omitted when there are good reasons to believe that there is very few, or nothing, to discover. Of course, consideration of this *pattern discovery* step would change the significance of the preparatory steps 4 and 5 above in this chapter.

4.3. Exploration steps

The following are corpus exploration steps (the first two are performed automatically).

1) For all predicate nouns of the list L and all patterns of the set P we run a concordancer-like procedure which retrieves fragments of texts matching the patterns (for performing corpus search operations the text processing toolkit called *UAM Text Tools* was used, cf. (Obrębski&Stolarski, 2006), this volume). This step is useful for two reasons: it permits finding the real text occurrences of the already identified collocations (those which belongs to the Basic Resource) but also to discovering the new ones, not attested yet. This second reason is the most interesting for us in this project.

2) The text fragments obtained within the step 1 are clustered with respect the predicate noun. In order to ease further manual processing, we have extracted from these fragments the entire list of these verbs that match the variable <verb> of the patterns (SVC). These are "candidates" to be identified as support verbs for the respective predicate nouns.

3) The list of support-verbs-candidates (SVC) is now to be processed (cleaned) manually, in order to eliminate parasite selections (large majority). This step requires investigating those text fragments where the candidates have been extracted from.

This last, third step is critical from the point of view of processing costs as the manual inspection is very slow (time-consuming).

4.4. Evaluation of the BR-enrichment algorithm

We have compared four different ways (henceforth we call them *variants*) to apply the procedures described in Chapters 4.2. and 4.3. These four variants differ according the set of patterns which is used to perform the exploration step 1. The preparatory steps are the same for these variants and the figures obtained on their application are as follows.

Basic Resource items: 2878
Corpus size: 80 000 000
List of BP predicate nouns (L): 2878 words
List of BP predicate nouns attested in the corpus:2406
Collocations in the Basic Resource (K): 5123
Structural schemes (S): 5404
Patterns (P): 257/56

Variant 1. We take as the set of patterns the general schemes obtained within the preparatory step 5c. These schemes are very general and are not likely to eliminate any good support-verb-candidate. What is to be expected is *overgeneration* of such candidates.

Variant 2. We consider as patterns which are to be used in the exploration step 1 the ones that are generated by the preparatory step 5b. Application of this set of patterns means that we restrict the search for support candidates to those verbs that may occur in the structural contexts of *some* collocations observed in the Basic Resource. Without a proof of *completeness*[3] of this set of patterns it must be assumed that there is a risk of limiting the *discovery potential* of the method.

Variants 1a and 2a. If the lists of the very support candidates obtained in the above two variants appear too long for further (hand) processing, it is possible to proceed as described in 4.2 and 4.3 with an additional limitation to consider only candidates among those already recognized as support verbs for *some* of predicate nouns of the Basic Resource (as well as their aspectual variants). For the respective two variants (denoted 1a and 2a) the set of the support candidates will be the same (the list of support verbs observed in the Basic Resource).

An evaluation of these variants was effected for a sample of 312 predicate nouns from the BP list (words beginning with the letters a, b or c).

The following were the main observed facts.

Variant 1
No. of patterns: 56
No. of text fragments matching the patterns[4]: 55929
No. of collocation candidates: 13579
No. of support verbs candidates: 2680

Variant 1a
No. of patterns: 56
No. of text fragments matching the patterns[5]: 32938
No. of collocation candidates: 5009
No. of support verbs candidates: 672

Variant 2
No. of patterns: 245
No. of text fragments matching the patterns[6]: 51649
No. of collocation candidates: 12462
No. of support verbs candidates: 2587

Variant 2a
No. of patterns: 245
No. of text fragments matching the patterns[7]: 30307

[3] By *completeness* we mean that the set of patterns covers all relevant cases. The completeness of the BP resource is to be a subject of further studies. Our preliminary, informal observations allow us to suppose that the completeness *degree* is rather high.
[4] Some of them may appear several times in the corpus
[5] Some of them may appear several times in the corpus
[6] Some of them may appear several times in the corpus
[7] Some of them may appear several times in the corpus

No. of collocation candidates: 4608
No. of support verbs candidates: 656

Comparing figures given above tell us that number of support candidates found when restricting search to the predefined list of BP support verbs almost equals the size of this list. Our claim is that limiting search to this list will not limit the substantially the discovery capacities of the method. On the other hand, the list of collocation candidates is much smaller for Variant 1a then for Variant1. The same holds for Variant 2 with respect to the Variant 2a. At the same time, the gain when passing from the more general form of patterns (variants 1 and 1a) to the less restrictive (variants 2 and 2a) will not substantially reduce the amount of manual work at step 3.

This means that for the whole lexicon we may expect ca 40 000-50 000 collocation candidates to be inspected at the exploration step 3. The size of this last data (SVC) directly determines the amount of manual processing necessary in order to complete the task.

Some additional processing may be considered in order to make the list of collocation-candidates shorter. The measure consists in systematic identification of parasite words which appear at the candidate lists, especially those with relatively high frequencies, and which for some reasons can not assume support functions. As example we may provide "jeść" which appears frequently at the candidate list as a result of ambiguous lemmatisation (the form "je" may be considered as a form of the verb "jeść" ("to eat") but also as a form of the personal pronoun "ona" ("she"). We are now compiling such a negative list that could be used to filter out parasite candidates.

4.5. Examples

The following example will present results of application of procedures described in 4.2 and 4.3 to a typical predicate noun described by (Vetulani, G., 2000). This is the predicate noun "bójka" considered in the BR as forming collocations with the support verbs "mieć" ("to have") and "wszcząć". Its formal description is:

bójka, f / mieć(Acc) / N1 z(Instr) / N2 o(Acc), wszcząć (Acc) / N1 z (Instr) / 2 o(Acc)

Variant 1 will generate the list "być, wybuchnąć, mówić, graniczyć, dojść, dochodzić, zwyknąć, zdarzać, wyniknąć, wspominać, **wdać**, trafić, **toczyć**, różnić, **rozpoczynać, rozpocząć**, powstać, pieprzyć, obyć, narzekać, **mieć**, miąć, **kończyć**".

Variant 2 will generate the list "być, wybuchnąć, dojść, dochodzić, zwyknąć, zdarzać, wyniknąć, wspominać, **wdać**, trafić, **toczyć**, **rozpoczynać, rozpocząć**, powstać, pieprzyć, narzekać, **mieć**, **kończyć**, graniczyć".

Variant 1a will generate the list "być, wychnąć, mówić, dojść, dochodzić, zwyknąć, **wdać**, **toczyć**, **rozpoczynać, rozpocząć**, **mieć**".

Variant 2a will generate the list "być, wychnąć, dojść, dochodzić, zwyknąć, **wdać**, **toczyć**, **rozpoczynać, rozpocząć**, **mieć**".

We can see that the predicate verb "wszcząć" (attested as support in the (Vetulani, G., 2000)) is not being retrieved. It is so because neither "wszcząć bójkę" nor "wszczynać bójkę" is not represented in the corpus. This example shows also a possible inconsequence of the Basic Resource, as "kończyć" is not retained as a support verb by the BR while "rozpoczynać/rozpocząć" are.[8]

The procedure we have described above has limitation typical of the corpus-based methods. One of them is the coverage problem. An example of this problem is the noun "bohomaz" which is considered by Vetulani (Vetulani, G. 2000) as predicative noun supported only by the verb "tworzyć" ("to create"). But the collocation "tworzyć bohomaz" is not observed in the corpus, and therefore none of the four variants may extract it. On the other hand, Variant 1 and Variant 2 propose as a collocation-candidate "

5. Conclusions

We pretend having demonstrated practical feasibility of machine assisted, corpus-based research in order to improve the dictionary-research-based dictionary of Polish predicate words. Full implementation of the described methods is our direct objective within the whole project of constructing Lexicon Grammar for Polish.

6. Acknowledgements

We wish to thank the Institute of Computer Science of the Polish Academy of Sciences, Warsaw, for providing us with the *IPI PAN Corpus* (http://www.korpus.pl) necessary for this research.

7. References

Jędrzejko, E. (1998). *Słownik polskich zwrotów werbo-nominalnych. Zeszyt próbny.* (in Polish), Energeia. Warszawa.

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

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

Vetulani, Z. (2000). Electronic Language Resources for POLISH: POLEX, CEGLEX and GRAMLEX. In M. Gavrilidou et al. (eds.), *Second International Conference on Language Resources and Evaluation, Athens, Greece, 30.05.-2.06.2000, (Proceedings),* ELRA, pp. 367-374.

Vetulani,, Z., Martinek, J., Obrębski, T. (2000). Dictionary-based tools for linguistic data acquisition from texts. In B. Lewandowska-Tomaszczyk, P.J. Melia (eds.), *PALC'99: Practical Applications in Language Corpora*, Peter Lang GmBH, Frankfurt, 2000, pp. 87-104.

Obrebski, T., Stolarski, M. (2006), "UAM Text Tools - a flexible NLP architecture", to appear in the Proceedings of LREC 2006.

[8] "Kończyć" ("to stop") and "rozpoczynać/rozpocząć" ("to start") are semantically and distributionally very close to each other. We can expect to have both of them being distributionally equivallent and we may expect them to be or both absent or both of them present in BR.