Wordnet Based Lexicon Grammar for Polish

Zygmunt Vetulani
Adam Mickiewicz University in Poznań, Poland
Department of Computer Linguistics and Artificial Intelligence
ul. Umultowska 87, 61-614 Poznań, Poland
E-mail: vetulani@amu.edu.pl

Abstract

In the paper we present a progress report on a long-term ongoing project concerning the lexicon-grammar of Polish. It is based on our former research focused mainly on morphological dictionaries, text understanding and related tools. By *Lexicon Grammar*s we mean grammatical formalisms which are based on the idea that a sentence is the fundamental unit of meaning and that grammatical information should be closely related to words. Organization of the grammatical knowledge into a lexicon results in a powerful NLP tool, particularly well suited to support heuristic parsing. The project is inspired by the achievements of Maurice Gross, Kazimierz Polanski and George Miller. We present the actual state of the project of a wordnet-like lexical network PolNet with particular emphasis on its verbal component, now being converted into the kernel of a lexicon grammar for Polish. We present various aspects of PolNet development and validation within the POLINT-112-SMS project.

Keywords: lexicon-grammar, lexical databases, collocations

1. Introduction

Already in the 1980s Antonio Zampolli argued for importance of language resources for future language industries. This visionary approach resulted in creation of institutions like ELRA/ELDA, conferences like LREC and, what is the most important, in global consensus about the necessity of development of language technologies for future growth. For some languages, like English, Italian or French an enormous amount of work has already been accomplished. On the other hand, there are still important gaps for most languages. For these languages, including Polish, the time is now to complete gaps in basic language resources and technologies, and this is to be done on national level. The long-term, on-going program we present here has been inspired mainly by three past reference project: Lexicon-Grammar (M. Gross, since the 70s), Syntactic-Generative Dictionary of Polish Verbs (K. Polański, 70s), Princeton WordNet (G.A. Miller, since the 90s).

The present project of a lexicon-grammar of Polish is a continuation of our earlier research focusing mainly on morphological dictionaries and related tools (presented at LREC 2000); cf. Z. Vetulani, (2000). The main achievement of this early stage was the morphological electronic dictionary POLEX, now publically accessible in the source form through ELDA. It contains over 100,000 entries in a both human and computer friendly format (Vetulani et al., 1998). It is to be emphasized that a good morphological dictionary with complete inflectional information is very important for language engineering purposes for highly inflectional languages like Polish or all other Slavonic languages.

2. The Lexicon-Grammar and syntactic description formalisms

What is important for us in the general idea of *Lexicon Grammar* is that this is a grammatical formalism based on the hypothesis that elementary sentence is the fundamental unit of meaning and the idea that the natural way to construct a grammatical lexicon is to directly link words with possibly complete grammatical information describing their syntactic and semantic properties . Verbs and other predicative words were first investigated and are of the main interest. The idea of lexicon grammar was inspired by Z.S.Harris' transformation theory and has been systematically developed since the early 1970s by Maurice Gross who implemented it in form of *syntactic tables* (initially for French, then for other languages). Within this approach, predicative words were studied from the point of view of their aptitude to form elementary sentences. At about the same time a Polish linguist Kazimierz Polański started his works on application of the transformational-generative model for systematic description of Polish verbs. This experiment resulted with "Syntactic-generative Dictionary of Polish Verbs" (Polański, 1992) in five volumes published during (1980-1992) (its forerunner appeared already in 1976). The formalism proposed by Polański in his dictionary is not machine-readable so that syntactic and semantic information must be human preprocessed before its integration into the machine interpretable lexicon grammar.

3. Initial PolNet

By "initial PolNet" we mean a lexical data base system of the type of Princeton WordNet built from scratch for Polish nouns (following the so called "merge model"

methodology). The PolNet project started in 2006. The resource development algorithm (Vetulani et al., 2007) implies use of traditional dictionaries of Polish language as the linguistic information source as well as of the DEBVisDic platform as a development tool (Pala et al., 2007). DEBVisDic has been used for synsets generation and for editing hyponymy/hyperonymy relations which are the basic relations organizing the noun part of PolNet. The project started with creation of synsets for nouns in an incremental way i.e. starting with general and frequently used vocabulary. More precisely, we selected the most frequent words found in a reference corpus of Polish (IPI PAN Corpus)[1] with one important exception made for methodological reasons. The reason was that we assumed possibly early validation of the resource in a real-size application for which an application-complete vocabulary was necessary. (Cf. the section on PolNet validation context below.) The initial PolNet was basically made of synsets built form simple (one word) nouns. This resource amounts now to some 11,700 synsets for over 20,300 word-senses (and 12,000 nouns). (The estimation of the effort invested in the development of the initial PolNet (for nouns) is 11 man-months of effective work.) At Figure 1 we present as an example of PolNet entries a (simplified) description of the synset composed of Polish synonyms of "school" in its traditional meaning , i.e. the synset {szkoła:1, buda:5, szkółka:1,...}.

```
<SYNSET>
<ID>PL_PK-518264818</ID>
<POS>n</POS>
<DEF>instytucja zajmująca się kształceniem; educational institution </DEF>
<SYNONYM>
  <LITERAL lnote="U1" sense="1">szkoła</LITERAL>    % szkoła=school
  <LITERAL lnote="U1" sense="5">buda</LITERAL>
  <LITERAL lnote="U1" sense="1">szkółka</LITERAL>
  .....
</SYNONYM>
<USAGE>Skończyć szkołę</USAGE>
<USAGE>Kierownik szkoły</USAGE>
.....
```

Figure 1: PolNet entry for "school"

4. Extension of (initial) PolNet to a lexicon-grammar

Extension of the initial PolNet to other kinds of grammatical categories is an important step towards a laxicon grammar for Polish. The first move in this direction was creation of a valency dictionary of verbs on the ground of existing linguistic knowledge and grammatical resources (Vetulani et al., 2010). The valency information (both syntactic and semantic) for verbs was then compiled into the PolNet format and integrated with the initial PolNet using the DEBVisDic platform (Horak et al., 2008).[2]

Lexical units, and more precisely verb+meaning pairs (verb word senses) were grouped into synsets on the basis of the relation of synonymy. In contrast to nouns, for which the interest is mainly in the hierarchical relations between concepts (represented by synsets) (hyperonymy/hyponymy), for verbs the main interest is in relating verbal synsets (representing predicative concepts) to noun synsets (representing general concepts) in order to show what are the semantic connectivity constraints corresponding to the particular argument positions. Inclusion of this information (combined with morphosyntactic constraints) gives to PolNet the status of a lexicon grammar.

This approach imposes granularity restrictions on verbal synsets and more exactly on the synonymy relation. We say that only such verb+meaning pairs are *synonymous* in which the same *semantic roles* take as a value the same concepts (this condition is necessary but not sufficient). In particular, the valency structure of a verb is one of formal indices of the meaning (so, all members of a given sysnset share the valency structure). This permits formal encoding of valency structure as a property of a synset.

Semantic roles as relations connecting noun synsets to verb synsets allow us to consider the extended PolNet as a *situational semantics network* of concepts. Indeed, as it is often admitted, verb synsets may be considered as representing situations (events, states), whereas semantic roles (Agent, Patient, Beneficent,...) provide information on the ontological nature of various actors participating, actively or passively, in this situation (event, state). Abstract roles (Manner, Time,...) refer to concepts which

[1] For this research we disposed of a 80 million words fragment (non-annotated) of the IPI PAN Corpus (http://korpus.pl/index.php?lang=en; access Mars 18, 2012). (At the time of application the whole corpus contained 200 million traditional words. Nowadays, it contains some 1,5 billion words and it is known as National Corpus of Polish (NKJP); cf. (Przepiórkowski et al. 2011)).

[2] http://deb.fi.muni.cz/clients-debvisdic.php (last access: Mars 18, 2012).

set the position of situation (event, state) in time, space and possibly also with respect to some abstract, qualitative landmarks. Formally, the semantic roles are functions (in mathematical sense) associated to the argument positions in the syntactic pattern(s) corresponding to synsets. Values of these functions are ontological concepts (here in form of noun synsets)[3]. E.g., for many verbs, the semantic role BENEFICENT takes as its value the concept representing the set of all humans (which are then considered as potential addressee of the situation effects).

```
<SYNSET>
<VALENCY>
<FRAME>Agent(N)_Benef(D)</FRAME>
<FRAME>Agent(N)_Benef(D) Action('w'+L)</FRAME>
<FRAME>Agent(N)_Benef(D) Manner</FRAME>
<FRAME>Agent(N)_Benef(D) Action('w'+L) Manner</FRAME>
</VALENCY>
<ILR type="category_domain" link="1356">CITTA:1</ILR>
<ILR type="Agent" link="ENG20-02383992-n">człek:1, człowiek:1, istota ludzka:1, zwierzę:2, ....</ILR>
<ILR type="Benef" link="ENG20-02383992-n">człek:1, człowiek:1, istota ludzka:1, zwierzę:2, ....</ILR>
<ILR type="Action" link="PL_PK-2035015933">czynność:1</ILR>
<ILR type="Manner" link="2214">CECHA_ADVERB_JAKOŚĆ:1</ILR>
<DEF>"wziąć (brać) udział w pracy jakiejś osoby (zwykle razem z nią), aby ułatwić jej tę pracę"</DEF>
<SYNONYM>
<WORD>pomóc</WORD>
<WORD>pomagać</WORD>
<LITERAL lnote="U1" sense="1">pomóc</LITERAL>
<LITERAL lnote="U1" sense="1">pomagać</LITERAL>
</SYNONYM>
<ID>3441</ID>
<USAGE>Agent(N)_Benef(D); "Pomogłam jej."</USAGE>
<USAGE>Agent(N)_Benef(D) Action('w'+L); "Pomogłam jej w robieniu lekcji."</USAGE>
<USAGE>Agent(N)_Benef(D) Manner Action('w'+L); "Chętnie pomogłam jej w lekcjach."</USAGE>
<USAGE>Agent(N)_Benef(D) Manner;"Chętnie jej pomagałam."</USAGE>
<CREATED>agav 2010-11-27 18:49:47</CREATED>
<POS>v</POS>
</SYNSET>
```

Figure 2: Description of the verbal synset {pomóc:1, pomagać:1} /to help/

Figure 2 presents a simplified description of the verbal synset {pomóc:1, pomagać:1} /to help/ in PolNet as it is displayed in DEBVisDic (with some slight modifications made for transparency) in the XML format.

In January 2012 the verb part of PolNet was composed of over 1,500 synsets corresponding to some 2,900 word+meaning pairs for 900 most important Polish simple verbs.

[3] We use PolNet synsets as role values, but it is possibly to use concepts from some general ontology, as e.g. Sumo, cf. (Pease, 2011).

5. PolNet validation context

The initial PolNet was tested and validated in a large-scale public security project (POLINT-112-SMS) with emulated language competence (Polish) (Vetulani, Z. et al. 2010a; Vetulani & Marciniak, 2011). The system POLINT-112-SMS understands and generates SMS messages. As the PolNet plays an essential role in the system we had to additionally feed the resource with the vocabulary typical of the system prototype application area (security at the football stadium)[4]. This additional vocabulary was extracted from small corpora (law text on public security issues /articles, legislation and books/, texts extracted and transcribed from emergency telephone[5] records (24 h of continuous speech records, 59.000 words))[6].

[4] The target applications will require further extensions.

[5] The Polish emergency telephone police service 997, now integrated into the 112 service.

[6] To collect additional information necessary to correct implementation of the POLINT-112-SMS prototype we used also two SMS dialogue corpora: a private SMS corpus of over 1700 messages (collected by Walkowska) and an SMS corpus (over 1000 messages) collected in an experimental setting of a role playing game (Vetulani et al. 2010a; Walkowska, 2009).

Within the POLINT-112-SMS system, PolNet is used for various tasks: for identification of collocations, for semantic interpretation of tokens and representation of sentences, for anaphora resolution and for disambiguation. PolNet is integrated with the system as one of modules of the POLINT-112-SMS architecture. It is interfaced by an access layer which execute queries written in the WQuery language (Kubis, 2011). WQuery language has also been used to verify the well-foundedness of PolNet, in particular in order to detect circular paths or synsets without hyperonyms. This method helped us to eliminate in a computer assisted way a number of wordnet construction errors which occurred at the encoding stage.

6. On-going and future work

The main concern of the project now is further extension of the verb PolNet to verb-noun collocations. This (current) work is based on the substantial preparatory steps started already in late 1990s by Grażyna Vetulani (Vetulani, G., 2000, 2012). This research consisted at first in manual examination of some 40,000 of Polish nouns in order to fix a list of over 7,500 abstract nouns playing the role of sentence predicate when supported with a semantically "empty" verb (light verb). From this set over 2,800 predicative nouns were subject of detailed description in terms of valency structure of the predicative constructions they form together with appropriate *light verbs*. This work was based initially on dictionary research (Vetulani, G., 2000), followed by further study involving corpus exploration (Vetulani Z. et al., 2007) (the IPI PAN corpus of Polish texts was used). As a result of this corpus-based step the number of identified and described collocations raised up from over 5,400 (for the dictionary-based step) to some 16,000.

The current step towards the Polish Lexicon Grammar (Vetulani, Z, Vetulani, G.; in print) consists in the incorporation of the noun-verb collocation lexicon to the PolNet, as we did before for simple verbs. Two important cases are considered, as they require different processing:

Case 1 (simple): the considered verb+noun collocation has a synonym[7] already included in PolNet. In this case we only need to enlarge the already existing synset.

Case 2 (more complex): the considered verb+noun collocation *does not have* a synonym in PolNet.

Recently, a Polish Government project (headed by G. Vetulani) has been opened in order to finalize the full integration of the already gathered verb-nouns collocation with the lexicon grammar in form of PolNet. (Cf. the Credits section below.) The achievement of this task is planned for 2013.

7. Availability

The version v0 of PolNet was first public released on November 25, 2011 at LTC 2011 (see www.ltc.amu.edu.pl). Both the initial PolNet and its verbal extension is now publically available for research purposes (free of charge). This work is licensed under a Creative Commons Attribution - Non Commercial - No Derivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd /3.0/; access on Mars 18, 2012).[8]

8. Credits

The research presented in this paper was partially covered by the Polish Government research grant R00 028 02 "Text processing technologies for Polish in application for public security purposes" (2006-2010) within the Polish Platform for Homeland Security and is continued under the MNiSzW grant 11H11 010080 "Lexicon-grammar oriented extension of Polish digital valency dictionaries for computer applications in humanities"(2012-2015).

9. References

Gross, M. (*1994).* Constructing Lexicon-Grammars, In *Computational Approaches to the Lexicon*, Oxford, Oxford University Press, pp. 213--263.

Horak, A., Pala, K., Rambousek, A. (2008). The Global WordNet Grid Software Design. In

Tanács, A. Csendes, D., Vincze, V., Fellbaum, C., Vossen, P. (Eds.), *Proceedings of the Fourth Global WordNet Conference*. Szeged: University of Szeged, pp. 194--199.

Kubis, M. (2011). An Access Layer to PolNet - Polish WordNet. InVetulani, Z. (Ed.), *Human Language Technology. Challenges for Computer Science and Linguistics. LTC 2009. Revised Selected Papers.* LNAI 6562, Springer-Verlag Berlin Heidelberg, pp. 444--455

Miller, G.M. (1995). WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11, pp.39--41.

Pease, A. (2011). *Ontology. A practical Guide.* Articulate Software Press, Angwin CA, USA.

Polański, K. (Ed.) (1992). *Słownik syntaktyczno - generatywny czasowników polskich.* Vol. I-IV, Wrocław: Ossolineum, 1980-1990, vol. V, Kraków: IJP PAN (1992).

Przepiórkowski, A., Bańko, M., Górski, R., Lewandowska-Tomaszczyk, B., Łaziński, M., Pęzik (2011). National Corpus of Polish. In Vetulani, Z. (Ed.), *Proceedings of the 5th Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, November 25-27, 2011, Poznań, Poland*, Wyd. Fundacja UAM, Poznań, pp. 259--263.

Vetulani, G. (2000). *Rzeczowniki predykatywne języka polskiego. W kierunku syntaktycznego słownika rzeczowników predykatywnych*. (In Polish). Poznań: Wyd. Nauk. UAM.

Vetulani, G. (2012). *Kolokacje werbo-nominalne jako*

[7]We must use the same synonymy relation as for one-word verbs, already considered in PolNet.

[8]For more information mail to: vetulani@amu.edu.pl.

samodzielne jednostki języka. Syntaktyczny słownik kolokacji werbo-nominalnych języka polskiego na potrzeby zastosowań informatycznych. Część I. (In Polish). Poznań: Wyd. Nauk. UAM.

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

Vetulani, Z., Dąbrowski, A., Obrębski, T., Osiński, J., Kubacki, P., Kubis, M., Marciniak, J., Vetulani, G., Walkowska, J., Witalewski, K. (2010a). *Zasoby językowe i technologie przetwarzania tekstu. POLINT-112-SMS jako przykład aplikacji z zakresu bezpieczeństwa publicznego*. (In Polish). Adam Mickiewicz University Press, Poznań.

Vetulani, Z., Kubis,M., Obrębski, T. (2010b). PolNet – Polish WordNet: Data and Tools. In Calzolari, N. (Ed.), *Proceedings of the seventh International conference on Language Resources and Evaluation (LREC 2010), May 19-21, Valletta, Malta, (Proceedings*), ELRA, Paris.

Vetulani, Z., Marciniak, J. (2011). Natural Language Based Communication between Human Users and the Emergency Center: POLINT-112-SMS. In Vetulani, Z. (Ed.), *Human Language Technology. Challenges for Computer Science and Linguistics. LTC 2009. Revised Selected Papers*. LNAI 6562, Springer-Verlag Berlin Heidelberg, pp. 303--314.

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 (2007), Menlo Park, California, pp. 267--268.

Vetulani, Z., Vetulani, G. (in print): Through Wordnet to Lexicon Grammar, in: Fryni Doa et al. (eds.) *Proceedings of the 30th Conference on Lexis and Grammar, 5-8.10.2011, Nicosia, Cyprus*, Éditions Honoré Champion.

Vetulani, Z., Walczak, B., Obrębski, T., Vetulani, G. (1998). *Unambiguous coding of the inflection of Polish nouns and its application in the electronic dictionaries - format POLEX*, Adam Mickiewicz University Press, Poznań.

Walkowska, J. (2009). Gathering and Analysis of a Corpus of Polish SMS Dialogues. In Kłopotek,M., Przepiórkowski, A., Wierzchoń, S. T., Trojanowski, K. (Eds.), *Challenging Problems of Science. Computer Science. Recent Advances in Intelligent Information Systems.* Academic Publishing House EXIT, Warsaw, pp. 145--157.