“PolNet - Polish WordNet” project: PolNet 2.0 - a short description of the release

Zygmunt Vetulani
Adam Mickiewicz University
Poznań, Poland
vetulani@amu.edu.pl

Bartłomiej Kochanowski
Adam Mickiewicz University
Poznań, Poland
bartlomiej.kochanowski@amu.edu.pl

Abstract

In December 2011/January 2012 we have released the main deliverable of the project "PolNet - Polish WordNet". It was first presented and distributed (as PolNet 1.0) at the 5th Language and Technology Conference in Poznań (2011) and (informally, with kind permission of the organizers) distributed during the Global Wordnet Conference in Matsue, Japan, in January 2012. We intend to present to the participants of the GWC 2014 the characteristics of the new, extended release of PolNet.

1 Introduction

In 1985 G. Miller with collaborators at the Princeton University initiated a novel method of systematizing semantic grammatical knowledge on the basis of the concepts of synonymy and hyperonymy. He proposed to organize a lexicon in the form of a lexical database (WordNet): a hierarchical network of a set of synonyms. The project appeared to be *generic* and inspired many followers working for various languages. Its practical value was recognized by language industries and practical computer science. In particular, lexical bases similar to Princeton WordNet (PWN) were used as ontologies useful in the AI oriented research.

2 Lexicon-grammar, VerbNet, FrameNet

The initial WordNet was organized as a set of equivalence classes with respect to the synonymy relation. For these classes, called *synsets*, other relations were considered, like hypernymy, meronymy, holonymy etc. Within the initial approach focusing on the meaning of words, only root forms of words were stored with no morphological or morphosyntactical information. Bringing this kind of information to wordnet is an idea which has as its forerunner the *lexicon-grammar* approach developed since the early 1970s (until late 1990s) by Maurice Gross (Gross, 1994) inspired by the works of Zellig S. Harris. Gross considered *elementary sentence* as a “minimal unit of sense” and the sense of a word as determined by the minimal sentences containing this word. This led to the concept of syntactic lexicon where grammatical information (syntactic) is contained in the lexical entries (in form of syntactic and semantic requirements /valences/ of predicative words). At about the same period (1980-1992), the similar ideas of Polański led to the monumental description of Polish verbs ("Syntactic-generative Dictionary of Polish Verbs" (Polański, 1992)). These works preceded (and perhaps even inspired) the future works in the FrameNet (Fillmore et al., 2002) and VerbNet (Palmer, 2009) projects which ware natural extensions of the initial WordNet. Independently from Polański, but following the same lines and applying refined Levis’ verb classes, Martha Palmers from the University of Colorado Boulder defined a lexical database where verbs were grouped according shared meaning and similar syntactic behavior (Palmer et al., 2005). These verb classes are “completely described by thematic roles, selectional restrictions on the arguments, and frames consisting of a syntactic description and semantic predicates with a temporal function”. VerbNet is sometimes compared to FrameNet, a kind of dictionary of word senses with annotated examples that show the meaning and usage. This project was initiated By Charles J. Fillmore in Berkeley (1997) and based on its concepts of frame semantics and semantic roles. Both VerbNet and FrameNet were applied in Artificial Intelligence (AI) projects concerning semantic processing of texts or machine question answering.

3 "PolNet - Polish WordNet" project

The project "PolNet - Polish WordNet" started in 2006. It was conceived in order to fill the technological gap consisting in the lack of a digital lexical database for the Polish.[1] The development algorithm (Vetulani et al., 2007) was based on several traditional dictionaries of the Polish language (in particular (Szymczak, 1978) and (Dubisz 2006)) and a general wordnet development tool which was the DEBVisDic platform (Pala et al, 2007).

The methodology we have applied to the development of PolNet followed the so called "merge model". PolNet was built from scratch involving intensive and large scale manual lexicographers' work. At the early stage of development we decided to abstain from any automatic synset generation and to reuse the existing knowledge about Polish accumulated by the past generations of linguists and lexicographers in lexicons, dictionaries and grammars.

The team - formed of computer scientists and lexicographers familiar with computer technologies explored first of all traditional resources (dictionaries). This work was inspired by and benefited from the methodology and tools of the EuroWordNet and Balkanet projects. In particular, production of synsets was supported by the VisDic and DEBVisDic systems generously made accessible for PolNet development by Karel Pala from the Masaryk University in Brno (Czech Rep.). PolNet 1.0 was made public available in November 2011. This first completed distribution was reduced to nouns and simple verbs (Vetulani and Obrębski, 2010).

The PolNet 1.0 release consisted of nominal and verbal synsets. Both the nominal and verbal parts were set up on the basis of frequency observed in the corpus (the IPI PAN corpus was used; cf. (Przepiórkowski, 2004)). The only systematic exception from this rule was made in order to be able to test PolNet in a real-scale application. This was the POLINT-112-SMS system, an application in the field of public security and PolNet, in which the latter one served as the ontology. It appeared necessary to extend the lexical coverage in the way to make PolNet complete with respect to the *a priori* chosen domain of public security at football stadiums. In the present development we continue on the ground of the frequent-concepts-first rule.

The noun part of the PolNet 1.0 consisted of the noun synsets partially ordered by the hyponymy/hyperonymy relation and the verb part was organized by the predicate-argument relationship connecting the verb synsets with the noun synsets. In the present extension (from PolNet 1.0 to PolNet 2.0) we will continue to apply this organization.

The main statistics of the PolNet 1.0 were as follows:

- Nouns: 11,700 synsets (12,000 nouns, 20,300 word+meaning pairs)
- Verbs: 1,500 synsets (900 verbs, 2,900 word+meaning pairs)

4 Extension motivations, reasons and policy

Although the usefulness of PolNet 1.0 as lexical ontology was confirmed through practical applications, we concluded the necessity of further extensions and improvements. The most fundamental decision was to consider as priority the development of the verbal component, before enlargement *ad infinitum* of the noun part. This decision was motivated by the practical needs of high quality, linguistically sound tools for advanced NLP, including text understanding, useful in Question Answering (QA), Machine Translation (MT) and other AI applications involving language competence modeling. We consider the extension to PolNet 2.0 described here as an important step towards a lexicon-grammar of Polish directly useful in systems development.

5 From PolNet 1.0 to PolNet 2.0

The present stage of the "PolNet - Polish WordNet" project consists of development from PolNet 1.0 to PolNet 2.0. The main task of this stage is to extend substantially the verbal component with the inclusion of concepts (synsets) represented (in many cases uniquely) by compound construction in form of verb-noun collocations (by verb-noun collocations we mean compound verbal structures made of a support verb and a predicative noun). This extension brought (until now) to PolNet some 1200 new

[1] PolNet shouldn't be confused with another wordnet for Polish (plWordNet) developed by Piasecki and others within a totally different methodology whose conception is based on automatic acquisition of synsets and relations.

verb synsets corresponding to 600 predicative nouns, some of those synsets being closely related to the already existing verb synsets of the PolNet 1.0.

The verb-noun collocation imported to PolNet come from the "Syntactic dictionary of verb-noun collocations in Polish" compiled by Grażyna Vetulani (Vetulani, G. 2000 and 2012).[2] See the Example 1 in Table 1 below.

Example 1. A fragment of the entry describing the predicative noun "pomoc" (compiled from a traditional dictionary):

pomoc, f/ *[help]*
udzielać(Gen)/N1(Dat),
"udzielać komuś pomocy" [to help](imperfective)
udzielić(Gen)/N1(Dat),
"udzielić komuś pomocy" [to help](perfective)
pospieszyć na(Acc)/N1(Dat)
"pospieszyć komuś na pomoc"
pospieszyć z(Instr)/N1(Dat)
"pospieszyć z pomocą ofierze wypadku"
[to help a victim]
przyjść z(Instr)/N1(Dat)
"przyjść z pomocą choremu"
[to help sb who is ill]
przyjść na(Acc)/N1(Dat)
"przyjść na pomoc oblężonemu miastu"
[to bring help to a surrounded town]

(N1(Dat) – complement in the dative case)

Table 1. Dictionary of verb-noun collocations (fragment)

Adding collocations to PolNet was not trivial because of specific syntactic phenomena related with collocations in Polish (systematic, although not general, change of syntactic requirements between the compound verb (verb-noun collocation) and its one-word synonym is required).

In PolNet, as in other wordnets, lexical units are grouped into synsets on the basis of the relation of synonymy. In opposition to nouns, where the interest is mainly in the hierarchical relations (hyperonymy/hyponymy) between concepts (represented by synsets) - 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 PolNet the status of a lexicon grammar. This approach imposes granularity restrictions on verbal synsets and more exactly on the synonymy relation.

Synonymous are solely such verb+meaning pairs in which the same semantic roles take as value the same concepts (this condition is necessary but not sufficient). In particular, the valency structure of a verb is one of the formal indices of the meaning (it is so that all members of a given sysnset share the valency structure). This permits to formal encoding of valency structure as a property of a synset.

Semantic roles as relations connecting noun synsets to verb synsets allow the extended PolNet to be considered 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 position the 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. The values of these functions are ontology concepts (here in form of noun synsets). 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 addresses of the situation effects).

In the project we use a well described set of semantic roles, adapted from works of Fillmore and later of Palmer (Fillmore 1977, Palmer 2009).

In the Example 2 in Table 2 below we may observe several inter-synsets relations which are used to express semantic requirements of the predicate (verb).

For example the “Semantic_role: [Action]” which connects the noun synset “{czynność:1}” *[activity]* to the verb synset “{pomóc:1, pomagać:1, udzielić pomocy:1, udzielać pomocy:1}” *[to help]*.tell us that the verb opens

[2] Over 14,300 collocations are described (and published) until now but this work is in progress.

an argument which must be filled by a term referring to some activity. Similarly, the relation “Semantic role [Benef]” indicates what kinds of entities may benefit of somebody’s assistance.

Example 2. DEBVisDic presentation of a PolNet synset containing both simple verbs and collocations(simplified):

POS: v ID: 3441

Synonyms: {pomóc:1, pomagać:1, **udzielić pomocy**:1, **udzielać pomocy:1**} (*to help*)

Definition: "wziąć (brać) udział w pracy jakiejś osoby (zwykle razem z nią), aby ułatwić jej tę pracę"
(*"to participate in sb's work in order to help him/her")*

VALENCY:

- Agent(N)_Benef(D)
- Agent(N)_Benef(D) Action('w'+NA(L))
- Agent(N)_Benef(D) Manner
- Agent(N)_Benef(D) Action('w'+NA(L)) Manner

Usage: Agent(N)_Benef(D); "Pomogłam jej." (*I helped her*)
Usage: Agent(N)_Benef(D) Action('w'+NA(L)); "Pomogłam jej w robieniu lekcji." (*I helped her in doing homework)*
Usage: Agent(N)_Benef(D) Manner Action('w'+NA(L));
"Chętnie udzieliłąm jej pomocy w lekcjach." (*I helped her willingly doing her homework*)
Usage: Agent(N)_Benef(D) Manner;
"Chętnie jej pomagałam." (*I used to help her willingly*)

Semantic_role: [Agent] {człek:1, człowiek:1, homo sapiens:1, istota ludzka:1, zwierzę:2, jednostka:1, łepek:3, łebek:3, łeb:5, głowa:8, osoba:1, twarz:2, umysł:2, dusza:3}
({man:1,...,animal:2,...})
Semantic_role: [Benef] {człek:1, człowiek:1, homo sapiens:1, istota ludzka:1, zwierzę:2, jednostka:1, łepek:3, łebek:3, łeb:5, głowa:8, osoba:1, twarz:2, umysł:2, dusza:3}
({man:1,...,animal:2,...})
Semantic_role: [Action] {czynność:1} (*{activity:1}*)
Semantic_role: [Manner] {CECHA_ADVERB_JAKOŚĆ:1} (*qualitative adverbial*)

Table 2. A PolNet 2.0 synset

6 Problems

In the case of Polish, our decision to make wordnet a type of lexicon-grammar through the inclusion of possibly all relevant grammatical information, appeared to be challenging in case of verb-noun collocations. This is because the traditional relation of synonymy is not invariant with respect to the syntactic requirements of predicative words. For example the simple word "nakarmić" and its synonym in form of the collocation "dać jeść" both correspond to the English "to feed". At the same time they do not have the same syntactic requirements, as "nakarmić" requires a complement in the accusative, whereas "dać jeść" - in the dative. Therefore, they should be put into different synsets of PolNet. This is because the synset of PolNet are intended to contain complete syntactic and semantic information about words, the same for all synset members.

In PolNet 2.0 we have applied the solution, which seems optimal from the practical (language engineering) point of view - to store them in separate synsets related by the transformational relation OBJECT_TRANS (ACC,DAT) which describes the difference of their syntactic properties.

7 Further research plans

“PolNet - Polish WordNet” project is in progress, and it will continue to be for the foreseeable future. The total number of verb-noun collocations is estimated to be largely more than 20 000 items. The set of 14,341 described until now was considered in order to select the most frequently used in texts and to include them in the first step of enlargement. We intend to continue this extension at least through 2014. In parallel to our present main priority, we continue work on further steps of the PolNet project in particular its alignment to the upper ontology SUMO, as well as on the extension of the net to more basic terms: nouns, verbs and collocations. The long term plan is to transform PolNet into a complete lexicon grammar of Polish integrating all grammatical information necessary (and sufficient) for AI and Language Engineering (LE) applications.

Acknowledgements

This work was done within the National Program for Humanities (grant 0022/FNiTP/H11/80/2011 managed by Grażyna Vetulani).

References

Stanisław Dubisz (ed.), 2006. *Uniwersalny słownik języka polskiego PWN*, (*Universal dictionary of Polish*, in Polish), 2nd edition, Wydawnictwo Naukowe PWN. Warszawa, Poland.

Maurice Gross, 1994. Constructing Lexicon-Grammars. In: Beryl T. Sue Atkins, Antonio Zampolli (eds.) *Computational Approaches to the Lexicon*, Oxford University Press. Oxford, UK, pp. 213–263.

Charles J. Fillmore, Collin F. Baker, Hiroaki Sato, 2002. The FrameNet Database and Software Tools. In: *Proceedings of the Third International Conference on Language Resources and Evaluation.* Vol. IV. LREC: Las Palmas.

Charles J. Fillmore, 1977. *The need for a frame semantics in linguistics. Statistical Methods in Linguistics*. Ed. Hans Karlgren. Scriptor

George A. Miller, 1995. WordNet: A Lexical Database for English. *Communications of the ACM* Vol. 38, (No. 11): 39–41.

Karel Pala, Ales Horák, Adam Rambousek, Zygmunt Vetulani, Paweł Konieczka, Jacek Marciniak, Tomasz Obrębski, Paweł Rzepecki , Justyna Walkowska, 2007. DEB Platform tools for effective development of WordNets in application to PolNet. In: Zygmunt 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, Wyd. Poznańskie. Poznań, Poland, pp. 514–518.

Martha Palmer, Paul Kingsbury, Dan Gildea, 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics* 31 (1): 71–106.

Martha Palmer, 2009. Semlink: Linking PropBank, VerbNet and FrameNet. In ***Proceedings of the*** *Generative Lexicon Conference*. Sept. 2009. GenLex: Pisa, Italy.

Kazimierz Polański (ed.), 1992. *Słownik syntaktyczno - generatywny czasowników polskich* vol. I-IV, Ossolineum, Wrocław,1980-1990, vol. V, Kraków: Instytut Języka Polskiego PAN.

Adam Przepiórkowski, 2004. *Korpus IPI PAN. Wersja wstępna (The IPI PAN CORPUS: Preliminary version).* IPI PAN, Warszawa, Poland.

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

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

Grażyna Vetulani, 2012. *Kolokacje werbo-nominalne jako samodzielne jednostki języka. Syntaktyczny słownik kolokacji werbo-nominalnych języka polskiego na potrzeby zastosowań informatycznych. Część I.* (In Polish). Wyd. Nauk. UAM. Poznań, Poland.

Zygmunt Vetulani, 2012. Wordnet Based Lexicon Grammar for Polish. In *Proceedings of the Eith International Conference on Language Resources and Evaluation (LREC 2012),* May 23-25, 2012. Istanbul, Turkey, (Proceedings), ELRA: Paris, France, pp. 1645–1649.

Zygmunt Vetulani, Tomasz Obrębski, 2010. Resources for Extending the PolNet-Polish WordNet with a Verbal Component. In: Bhattacharyya, Pushpak, Fellbaum, Christiane, Vossen, Piek (eds.) *Principles, Construction and Application of Multilingual Wordnets. Proceedings of the 5th Global Wordnet Conference.* Narosa Publishing House: New Delhi, Chennai, Mumbai, Kolkata, pp. 325–330.

Zygmunt Vetulani, Grażyna Vetulani (in print): Through Wordnet to Lexicon Grammar. In: Fryni Kakoyianni Doa (Ed.). *Penser le lexique-grammaire : perspectives actuelles,* Editions Honoré Champion. Paris, France.

Zygmunt Vetulani, Justyna Walkowska, Tomasz Obrębski, Paweł Konieczka, Paweł Rzepecki, Jacek Marciniak, 2007. PolNet - Polish WordNet project algorithm, 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, 2005,* Wyd. Poznańskie, Poznań, Poland, pp. 172–176.