Recent Advances in Development of a Lexicon-Grammar of Polish: PolNet 3.0

Zygmunt Vetulani, Grażyna Vetulani, Bartłomiej Kochanowski
Adam Mickiewicz University in Poznań
ul. Wieniawskiego 1, 61-712 Poznań, Poland
E-mail: vetulani@amu.edu.pl, gravet,@amu.edu.pl, bkochanowski@amu.edu.pl

Abstract

In this paper we present recent works contributing to transformation of the initial PolNet, a Polish wordnet developed at the Adam Mickiewicz University, into a Lexicon Grammar of Polish. We focus on granularity issues that occurred at the stage of including verb-noun collocations as well as information related to language registers.

Keywords: wordnet, synonimy, granularity, valency structure, lexicon grammar, collocations, registers

1. Introduction.

In the mid-1980s, G. Miller started the development of a novel approach to systematize the semantic description of words (Miller, 1985). The leading idea was to organize a lexicon as a lexical database (Princeton WordNet / PWN) consisting of a hierarchical network of classes of synonyms. PWN appeared useful in AI applications involving NLP. Some PWN followers decided to go one step further and enrich the word descriptions with complex data structures to represent events, relations and situations. The forerunners of this idea (Gross (1994) and Polański (1992)) considered the *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 *lexicon-grammar*[1] where grammatical information sufficient to describe elementary sentences is contained in the lexical entries and where the elementary sentence is the basic unit of meaning. Their contributions preceded the future works within the FrameNet (Fillmore et al., 2002) and VerbNet (Palmer, 2009) projects. The first one referred to Frame Semantics developed by Fillmore. Frames describe lexical units (typically verbs) and their syntactic dependents characterized by frame elements. In the VerbNet project, verbs are grouped according to shared meaning and similar syntactic behavior. Palmer used thematic roles, selectional restrictions on the arguments, and frames containing syntactic description of the verb.[2]

2. The initial PolNet: inspiration and methodology

We started PolNet in 2006 intending to build a lexical ontology as a wordnet similar to PWN. Initially PolNet was implemented for nouns. We decided to compile PolNet from scratch (*merge development model*[3]) in the way inspired by the PWN and EuroWordNet projects (Vossen 2002). This method guarantees (quasi)one-to-one correspondence between the structure of synsets and the conceptualization shared by a (quasi)totality of users of the language concerned[4]. This is an important quality factor often underestimated by the wordnet designers applying the less costly *expand model*. The first milestone (2009) of the PolNet project was reached on attaining over 10K synsets[5] for some 10K words (corresponding to almost 19K word+meaning pairs).[6] The selection of the lexical material for the initial PolNet was importance-driven. A major subset of nouns was taken from the frequency list (compiled for the IPI PAN Corpus (Przepiórkowski, 2004) and the list of semantic descriptors (761) used by Polański (Vetulani, Z., 2003) to express semantic restrictions on verb arguments[7]. The PolNet development algorithm (Vetulani, Z. et al., 2007) was based on several traditional dictionaries of Polish and the DEBVisDic platform (Pala et al. 2007). As a test-bed for using PolNet as an ontology we chose the Polint-112-SMS system (Vetulani and Marciniak, 2011) with natural language understanding functionality (homeland security domain). For testing purposes we augmented the lexical coverage by domain specific terminology.

3. Addition of the verbal component: from the initial PolNet to a lexicon-grammar

Our intention to make PolNet useful for systems with language functionality was the reason to extend PolNet to verbs (initially simple, then compound). This was also the first step to transform a lexical ontology for Polish (PolNet) into a lexicon-grammar (Vetulani, Z. Obrębski, T., and Vetulani, G. 2007). In a lexicon-grammar, to describe the

[1] First developed for French (since the early 1970s until late 1990s; Gross 1994). This idea was already implemented in our first implementations of the NL interfaces for Polish (Vetulani 1988). Several large scale projects have been recently launched in the area of valency dictionaries both for simple and compound verbs (Vetulani G. 2000, 2012), (Przepiórkowski et al. 2014).

[2] http://verbs.colorado.edu/~mpalmer/projects/verbnet.html

[3] For merge/expand models see (Vossen, 2002), p. 52, section 3.1.

[4] We are aware that some experts may contest this statement as too strong, but we bring the reader's attention to the fact that this is a matter of granularity (see section 4 below).

[5] In PolNet 3.0 (now) the number of synsets is 12,011 for nouns, and 3,645 for simple and compound verbs (corresponding to 28,574 word+meaning pairs).

[6] Some 2,400 of these synsets were aligned to the PWN equivalents.

[7] Some examples of semantic descriptors proposed by Polański (Vetulani, Z., 2003): instrument (instrument), organ (body part), zwierzę (animal), roślina (plant), kwiat (flower), drzewo (tree),…

meaning of a predicative word one seeks to characterize the set of elementary sentences having this word as predicate. Our method of implementing this idea differs from what Gross did for French (through syntactic tables). In particular we introduce the so called *semantic role relations* such as Agent, Object, Patient, Beneficient[8] between verb and noun synsets to encode how the verbs and nouns combine to form simple sentences.

POS: v ID: 3441
Synonyms: {pomóc:1, pomagać:1, **udzielić pomocy**:1, **udzielać pomocy:1**} (*to help*)
Definition: "wziąć udział w pracy jakiejś osoby, 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] {człowiek:1, homo sapiens:1, istota ludzka:1, …} (*{man:1,...,human being:1,...}*)
Semantic_role: [Benef] {człowiek:1, homo sapiens:1, istota ludzka:1, …} (*{man:1,...,human being:1,...}*)
Semantic_role: [Action] {czynność:1} (*{activity:1}*)
Semantic_role: [Manner] {CECHA_ADVERB_JAKOŚĆ:1} (*qualitative adverbial*)

Fig. 1. Simplified DEBVisDic presentation of a PolNet synset containing both simple verbs and collocations (Vetulani and Kochanowski, 2014).

The first stage of extending PolNet with 900 simple verbs carefully selected among the most important verbs was done in a relatively short time due to the high quality of the description of Polish verbs. This stage resulted in the publicly available release of PolNet 1.0 under a CC license and distributed at the LTC (November, 2011) and shortly after at the Global Wordnet Conference (January, 2012).[9]

4. Recent enlargement of PolNet: granularity and other issues

Extension of the initial PolNet with simple verbs (PolNet 1.0) and verb-noun collocations[10] (PolNet 2.0)[11] opened up new application opportunities and motivated us to re-consider the fundamental problem of synonymy, directly related to the granularity of the wordnet. For verbs, and generally for all predicative structures, we focus on relations between the *verb synsets* (corresponding to predicative concepts) and *noun synsets* (representing nominal concepts), rather than on hierarchical relations, in order to show the semantic/morpho-syntactic connectivity restrictions corresponding to arguments. For these words, we propose to refine the concept of synonymy by considering valency structures. By *valency structure* we mean *the structured information on the arguments opened by the predicative word including both semantic constraints on the arguments (semantic role values) as well as the surface morpho-syntactic and pragmatic properties of the text fillers of argument positions (case, number, gender, preposition, register etc.)*[12]. The valency structure of a verb is one of the formal indices of meaning and should be considered as an attribute of a synset, i.e. all synset's members should share the valency structure. Strict application of this principle results in a fine granularity of the verb section of the wordnet.[13]

Extending the initial PolNet (in particular adding collocations) was not straightforward because of specific phenomena frequent in highly inflected languages but rare in low inflected ones. Paraphrasing a sentence by replacing it's verb by a collocation often requires change of the argument's grammatical case. Although the simple verbs "kupić" ("to buy"), "nabyć" ("to buy"), as well as the collocation "dokonać zakupu" ("to make a purchase"). may all be translated into *to buy* in English, the grammatical case of the inanimate object ("towar"/"goods") will change from Accusative to Genitive when replacing any of the simple verbs (*kupić/nabyć towar(Acc)* by the collocation *dokonać zakupu towaru (Gen)*). To simplify further processing, we decided to apply our definition of synonymy rigorously. This decision implies storing collocations and their corresponding single word equivalents in separate synsets, if only their valency structures are different (even if the intuitive meaning and

[8] In PolNet we use a set of semantic roles adapted from Filmore (1977) and Palmer (2009).
[9] It is free available through www.ltc.amu.edu.pl and from Meta-Share.
[10] Inclusion of verb-noun collocation in a relatively short time was possible on the ground of earlier works (Vetulani 2000, 2012).
[11] Cf. (Vetulani and Kochanowski, 2014).
[12] Considering registers as distinctive for synsets is novel for wordnets and opens the pragmatic dimension. We apply the following registers in inspired by ISO 12620: neutral, dialect, formal, informal, ironic, register, taboo, technical, vulgar, archaic (not in ISO), literary (not in ISO).
[13] Information about the valency structure appeared very efficient in the heuristic, rule-based parsers where the valency was explored at the pre-analysis stage (Vetulani and Marciniak, 2011).

usage seem be identical). However, in all such cases we keep the corresponding synsets related by the transformational relations which describe the differences among their morpho-syntactic properties. Fig. 2. presents the (fragment of) valency structures /simplified/ for the verbs "kupić" and "nabyć" in opposition to the valency structure for "dokonać zakupu". We observe the grammatical case transformation of the direct object between a sentence and the collocation-based paraphrase.

"*Piotr kupił* ***mieszkanie(Acc)***"
<VALENCY>
<FRAME>Agent(N) _ **Object(Acc)** </FRAME>
</VALENCY>
"*Piotr nabył* ***mieszkanie(Acc)***"

<VALENCY>
<FRAME>Agent(N) _ **Object(Acc)**</FRAME>
</VALENCY>

"*Piotr dokonał zakupu* ***mieszkania(Gen)***"
<VALENCY>
<FRAME>Agent(N) _ **Object(D)**</FRAME>
</VALENCY>

Fig. 2. Case transformation of the Object

In PolNet we store simple verbs "kupić" and "nabyć" (for one of their possible common meanings) in the same synset, whereas the collocation "dokonać zakupu" is to be included in a different one. These synsets are related by an external (inter-synset) relation describing the direct object case transformation necessary for paraphrasing:

(TRANS_CASE_OBJECT(A,D)).

For the present refinement of PolNet 2.0, we have assumed that the category of language register is a part of the meaning. The totality of PolNet 2.0 synsets has been revised in order to split these PolNet 2.0 synsets that contain different register words into register-uniform sub-synsets. The initial synset is then retracted, and all the subsynsets (with identical valency structure except for register) are introduced instead and related by the relation of synset similarity. This procedure has been completed for 638 *basic synsets*, i.e. synsets that may serve to describe semantic properties of the argument positions opened by verbs resulting with 827 synsets.

	PolNet 0.1 (2009) [14]	PolNet 1.0 (2011)[15]	PolNet 2.0 (2013)	PolNet 3.0 (2016)
Nouns	10,629	11,700	11,700	12,011
Simple verbs	---	1,500	1,500	3,645
Collocations	---	---	1,200 [16]	1,908

Fig. 3. Growth of the PolNet's main parts (in synsets[17]). Notice. This table does not represent the effort invested in the development of PolNet as an important deal of work was engaged in the wordnet cleaning operations.

5. Future work

The version PolNet 3.0 which contains the recent improvements and extensions has already been user-tested as a resource for modeling semantic similarity between words (Kubis, 2015). We intend It will to proposed it for distribution through Data Centers (ELRA, META-SHARE) under a CC license. In the future, we plan both quantitative enlargement of the existing categories as well as inclusion the parts of speech not considered so far.

6. Credits

The recent results presented in this paper were mainly obtained within the Polish National Program for Humanities and were covered by the grant 0022/FniTP/H11/80/2011 (2012-2015). Earlier works on PolNet were partially covered by the Polish Government grant MNiSzW nr R00 028 02 (2006-2010), and the grant of the City of Poznań RoM.III/3420-52/10 Fn2625/10 (2011).

7. Bibliographical References

Gross.M., (1994). Constructing Lexicon-Grammars. In B.T.S.Atkins and A. Zampolli (Eds.). *Computational Approaches to the Lexicon*, Oxford University Press, UK, pp. 213–263.

Kubis, M. (2015). A semantic similarity measurement tool for WordNet-like databases. In Z. Vetulani and J. Mariani

[14] Vetulani, Z., Kubis,M., Obrębski, T. (2010). PolNet – Polish WordNet: Data and Tools, LREC 2010
[15] (Vetulani, 2014)
[16] (Vetulani and Kochanowski, 2014),
[17] The number of synsets should not be confused with the number of words.

(Eds), *Proceedings of the 7th Language and Technology Conference*, Poznań, Poland, 27-29 November 2015. FUAM, Poznań, pp. 150--154.

Fillmore, Ch., Baker, C.F. and Sato, H. (2002). The FrameNet Database and Software Tools. In *Proceedings of the Third International Conference on Language Resources and Evaluation*. Vol. IV. LREC: pp. Las Palmas, pp. 1157—1169.

Miller, G. A., Beckwith, R., Fellbaum, Ch, Gross, D. and Miller, K.. (1990). WordNet: An online lexical database. *Int. J. Lexicography.* 3, 4, pp. 235–244.

Pala, K., HORÁK, Horák, A., Rambousek, A., Vetulani, Z, Konieczka, P., Marciniak, J., Obrębski, T., Rzepecki, P. & Walkowska, J. (2007**).** DEB Platform tools for effective development of WordNets. In application to PolNet**.** In Z. Vetulani (Ed.). *Proceedings of the LTC 2007,* Wyd. Poznańskie, Poznań, pp. 514-518.

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

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

Przepiórkowski, A. (2004). *The IPI PAN Corpus. Preliminary Version.* Warszawa : IPI PAN.

Przepiórkowski, Hajnicz,E., Patejuk, A., Woliński, M., Skwarski, F., Świdziński, M. (2014). Walenty: Towards a comprehensive valence dictionary of Polish. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation.* LREC: Rejkyavik, pp. 2785-2792.

Vetulani, Z. (1988). PROLOG Implementation of an Access in Polish to a Data Base, In *Studia z automatyki*, vol. XII, PWN, 1988, pp. 5-23.

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

Vetulani, Z. (2003). Linguistically Motivated Ontological Systems, in: N. Callaos, W. Lesso, K.D. Schewe, E. Atlam, (Eds.). *Proceedings of the 7th World Multiconference on Systemics, Cybernetics and Informatics*, vol. XII, Int. Inst. of Informatics and Systemics, pp. 395-400.

Vetulani, Z., Walkowska, J., Obrębski, T., Konieczka, P., Rzepecki, P. and Marciniak,J. (2007). PolNet - Polish WordNet project algorithm. Z. Vetulani (Ed.). *Proceedings of the LTC 2007,*Wyd. Poznańskie, Poznań, pp. 172-176.

Vetulani, Z., Obrębski, T., Vetulani, G. (2007). Towards a Lexicon-Grammar of Polish: Extraction of Verbo-Nominal Collocations from Corpora. In *Proceedings of FLAIRS-07*, AAAI Press, Menlo Park, pp. 267-268.

Vetulani, Z. and Marciniak, J. (2011). Natural Language Based Communication between Human Users and the Emergency Center: POLINT-112-SMS. *LectureNotes in Artificial Intelligence* 6562, Springer-Verlag, pp. 303-314.

Vetulani, Z. and Vetulani, G. (2013). Through Wordnet to Lexicon Grammar. Fryni Kokoyianni-Doa (Ed.). *Penser le lexique-grammaire : perspectives actuelles,* Editions Honoré Champion, Paris, pp. 531-545.

Vetulani, Z., Kochanowski, B. (2014). PolNet – Polish WordNet project : PolNet 2.0 - a short description of the release , In: H. Orav, Ch. Fellbaum, and P. Vossen. (Eds.), *Proc. of the Seventh Global Wordnet Conference*, Jan. 2014, Tartu, pp. 400-406.

Vetulani, Z. (2014).PolNet-Polish Wordnet, In *Lecture Notes in Artificial Intelligence* 8387, Springer-Verlag, pp. 408-416.

Vossen, P., (Ed.) (2002). *EuroWordNet General Document, Version 3. Final, July 1.* http://www.vossen.info/docs/2002/EWNGeneral.pdf*;* access 12.11.2015).

13. Language Resource References

The last public release of PolNet – Polish Wordnet may be found at the ISRLN site through http://www.islrn.org/ using the resource ISLRN number 944-121-942-407-9.