PolNet – Polish WordNet

Zygmunt Vetulani[(✉)]

Department of Computer Linguistics and Artificial Intelligence,
Adam Mickiewicz University, Poznań, Poland
`vetulani@amu.edu.pl`

Abstract. We present a long term project aiming at the construction of a lexical database and ontology for Polish. The specific objective of the PolNet project is to provide a human-and-computer friendly description of the Polish language for direct application in language processing software.

Keywords: WordNet · FrameNet · VerbNet · PolNet · Wordnets · Linguistically motivated ontology · Collocations

1 Introduction

The "PolNet-Polish Wordnet" project started in 2006 within the grant of the Polish Platform for Homeland Security "Language Resources and text processing technologies oriented to public security applications". The project, inspired by the Princeton WordNet, intended to serve two main objectives: first to provide a reference semantic lexicon for the core of the Polish language, second to serve as a linguistically motivated ontology to support reasoning in the AI systems with natural language competence. The research was anticipated by our earlier studies on linguistically motivated ontologies (cf. Vetulani 2003, 2004). As is the case for practically all wordnet development projects, PolNet is a time-unlimited project run by an interdisciplinary team of experts involving linguists, language engineers and computer scientists[1]. Since the beginning, the project has evolved from a network of noun-based synsets[2] organized in a hyponymy/hyperonymy hierarchy towards a lexicon grammar containing a verbnet part as backbone.

2 Methodology

The initial PolNet was built as a wordnet for nouns organized into synsets interrelated first of all by relations corresponding to traditional hyponymy/hyperonymy. Keeping

---

[1] Since very beginning PolNet has been realized as a team project involving many people, each contributing with experience and enthusiasm. In this group were: Agnieszka Kaliska, Bartłomiej Kochanowski, Paweł Konieczka, Marek Kubis, Jacek Marciniak, Beata Nadzieja, Tomasz Obrębski, Paweł Rzepecki, Grzegorz Taberski, Agnieszka Vetulani, Grażyna Vetulani, Zygmunt Vetulani, Justyna Walkowska, Marta Witkowska, Weronika Wojciechowska. Several are still active in the project.

[2] By *synset* we mean a class of synonyms.

in mind that a correctly constructed wordnet should render the conceptualization of the real word (reflected in the language), we decided to apply the so called "merge development model" where the lexical network is built from scratch, in opposition to the so called "expand method" consisting in "translating" a wordnet built for some other language (typically English). Application of the merge model is expensive and time consuming, as it requires essential involvement of well trained staff, mainly lexicographers. Building a wordnet from scratch, we followed the methodology elaborated within the Princeton WordNet and EuroWordNet projects so that we could reuse the existing linguistic knowledge accumulated in traditional written grammatical resources, mainly dictionaries and grammars.

In order to limit methodological arbitrariness, we deliberately decided to avoid statistic and AI-based methods (like genetic algorithms, machine learning, etc.) at this stage.[3] The PolNet development algorithm (Vetulani et al. 2007) makes essential use of traditional Polish dictionaries and the DEBVisDic platform (Pala et al. 2007) as a wordnet development tool. The work was organized in an incremental way, starting with general and frequently used vocabulary extracted from a corpus (IPI PAN corpus; Przepiórkowski 2004). The one important exception to this rule was caused by the need to have the system tested in a real application at a possibly early stage of development. This is why we introduced, since the beginning, some basic domain terminology from the area of homeland security. (From 2009 to 2010 PolNet was tested as ontology in the application POLINT-112-SMS (Vetulani et al. 2009, 2010a, b)).

With inclusion of the verb category we brought ideas inspired by the FrameNet (Fillmore et al. 2002) and VerbNet (Palmer 2009) projects to PolNet. The verbal part became the backbone of the whole network. Its organizing part was the system of semantic roles.

Semantic roles, as relations connecting noun synsets to verb synsets, describe the semantic requirements of the predicate. This permits us to consider the verb-extended PolNet as a situational semantics network of concepts where verb synsets represent situations (events, states), whereas semantic roles (Agent, Patient, Beneficent,…) provide information on the ontological nature of the actors involved in the situation. The abstract roles (Manner, Time,…) describe the situation (event, state) with respect to time, space and possibly also to some abstract, qualitative landmarks. Formally, the semantic roles are functions (in a mathematical sense) associated to the argument positions in the syntactic patterns corresponding to synsets. Values of these functions are ontological concepts represented by synsets[4]. For example, for many verbs, the semantic role BENEFICENT takes as value the concept of *humans*. The set of semantic roles we used is adapted from (Palmer 2009).

---

[3] PolNet should not be confused with another wordnet project for Polish, i.e. "plWordnet" (Piasecki team, Wrocław) which uses intensively such methods in order to automatize (and speed-up) the synsets production process. Within the PolNet project, application of AI and statistical methods to accelerate the wordnet development will be considered at future stages, after having completed the hard core of the wordnet which will be used as the training resource for the sophisticated AI methods.

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

Example 1: entry for "school":

```
<SYNSET>
<ID>POL-2141601597</ID>
<POS>n</POS>
<DEF>Pies o umaszczeniu czarnym, podpalanym bądź brunatnym.</DEF
<SYNONYM>
<WORD>wilk</WORD>
<WORD>owczarek</WORD>
<WORD>niemiecki</WORD>
<WORD>owczarek</WORD>
<WORD>alzacki</WORD>
<WORD>wilczur</WORD>
<LITERAL lnote="U2a" sense="2">wilk</LITERAL>
<WORD>wilk</WORD>
<LITERAL lnote="U1" sense="1">owczarek niemiecki</LITERAL>
<WORD>owczarek</WORD>
<WORD>niemiecki</WORD>
<LITERAL lnote="U1" sense="1">owczarek alzacki</LITERAL>
<WORD>owczarek</WORD>
<WORD>alzacki</WORD>
<LITERAL lnote="U1" sense="1">wilczur</LITERAL>
<WORD>wilczur</WORD>
</SYNONYM>
<ILR type="hypernym" link="POL-2141701478">owczarek:1</ILR>
<BCS/>
<NL>false</NL>
<STAMP>krzyzang 2007-07-15 12:52:15</STAMP>
<CREATED>krzyzang 2007-07-15 12:52:15</CREATED>
</SYNSET>

<ILR type="hypernym" link="POL-2141701478">owczarek:1</ILR>
<BCS/>
<NL>false</NL>
<STAMP>krzyzang 2007-07-15 12:52:15</STAMP>
<CREATED>krzyzang 2007-07-15 12:52:15</CREATED>
</SYNSET>
```

3   Development Phases

The PolNet project was supported by several research programs and grants. After the initial phase of the noun-based lexical ontology development (to be used in the public security application), PolNet continued with the support of the City of Poznań[5], then within a resources-oriented project of the National Program for Humanities (NPRH)[6], and finally as a part of research program of the Faculty of Mathematics and Computer Science AMU. At the end of the first phase (March 2010) PolNet was built of env. 10600 synsets, 18800 word senses for 10900 nouns. At that time preparative works for inclusion of the verb component were already advanced. For this resource we did mapping experiments between PolNet and the so called Global Wordnet Grid (for 1200 noun synsets), as well between PolNet and Princeton Wordnet (for env. 2400 synsets). Intensive works within the CITTA-Ontology project resulted in the first public PolNet distribution at the Language and Technology Conference 2011 (November) in Poznań, Poland and the Global Wordnet Conference 2012 (January) in Matsue, Japan.

This first release was freely distributed as PolNet 1.0 under a CC (Creative Commons) license. The released resource amounted – for nouns – to approximately 11,700 synsets for over 20,300 word-senses (and 12,000 nouns). The verb part of PolNet was composed of more than 1,500 synsets corresponding to some 2,900 word + meaning pairs for 900 of the most important Polish simple verbs. We refer to this release as to PolNet 1.0.

4   From PolNet 1.0 to PolNet 2.0

The present development is being operated within the NPRH project "Development of Digital Resources of Polish in the Area of Valency Dictionaries Towards the Lexicon Grammar Oriented to Computer Applications in the Humanities"/11H11 010080/. It focuses on development of the verbnet part of PolNet. Although verb synsets may be related to other verb synsets through the hyponymy/hyperonymy relation, the main interest is in relating verb synsets (representing predicative concepts) to noun synsets (representing general concepts) in order to show the semantic connectivity constraints that correspond to the particular argument positions opened by verbs. Inclusion of the predicative information, combined with morphosyntactic constraints, gives PolNet the status (and strength) of a lexicon grammar. Our attempts to transform PolNet into a Lexicon Grammar of Polish were inspired by two historical reference projects of the 1970s: Lexicon-Grammar (Gross 1994) and Syntactic-Generative Dictionary of Polish Verbs (Polański 1992).

---

[5] "CITTA-ontology" to develop resources for tourism-oriented applications.

[6] "Development of digital resources of Polish in the area of valency dictionaries towards the lexicon grammar oriented to computer applications in the humanities" (Grant of the Polish Ministry of Research and Higher Education/11H11 010080).

At the present stage we focus our efforts on inserting verb-noun collocations into PolNet. Collocations assume predicative functions in a similar way to simple verbs. Verb-noun collocations often do not have one-word synonyms and, therefore, must be considered an essential part of the Polish predication system.

Example 2. 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" [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)

Our present work benefits from the recent advances of descriptive research on collocations. We directly use the "Dictionary of Polish Verb-Noun Collocations" (Vetulani, G. 2000, 2012) as basic resource. This resource is still in development and we may expect its essential expansion in the future. The most challenging and time consuming part of our work consists in adding semantic information missing in the G. Vetulani's verb-noun collocations dictionary (cf. Example 1 above for the form of dictionary entries). This part of work may hardly be done automatically without uncontrolled quality loss.

The intended release of PolNet 2.0 (planned for 2014)/2015 will contain new synsets corresponding to some 3600 collocations.

Example 3. 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 udzieliłam 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*)

5   Future Research

The "PolNet - Polish WordNet" project is in progress, and will be continued in the foreseeable future. Our short term priority is to complete first of all the verbnet part of the project. This means inclusion of both simple verbs and collocations. In parallel, we plan to continue our work, already started for PolNet 1.0, on the alignment of synsets to the upper ontology SUMO (Pease 2011). The long term goal doesn't change: it is to transform PolNet into a complete lexicon grammar of Polish integrating all grammatical information necessary (and sufficient) for advanced AI and Language Engineering (LE) applications.

Bibliography and References

PolNet Bibliography

Vetulani, Z., Walkowska, J., Obrębski, T., Konieczka, P., Rzepecki, P., Marciniak, J.: PolNet - Polish WordNet project algorithm. In: Vetulani, Z. (ed.) Proceedings of the 3rd Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, 5–7 October 2007, Poznań, Poland, pp. 172–176. Wyd. Poznańskie, Poznań (2007)

Pala, K., Horák, A., Rambousek, A., Vetulani, Z., Konieczka, P., Marciniak, J., Obrębski, T., Rzepecki, P., Walkowska, J.: DEB Platform tools for effective development of WordNets in application to PolNet. In: Vetulani, Z. (ed.) Proceedings of the 3rd Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, 5–7 October 2007, Poznań, Poland, pp. 514–518. Wyd. Poznańskie, Poznań (2007)

Vetulani, G., Vetulani, Z., Obrębski, T.: Verb-noun collocation SyntLex dictionary - corpus-based approach. In: Proceedings of 6th International Conference on Language Resources and Evaluation, 26 May–1 June 2008, Marrakech, Morocco (Proceedings). ELRA, Paris (2008)

Vetulani, Z., Walkowska, J., Obrębski, T., Marciniak, J., Konieczka, P., Rzepecki, P.: An algorithm for building lexical semantic network and its application to PolNet - Polish WordNet project. In: Vetulani, Z., Uszkoreit, H. (eds.) LTC 2007. LNCS (LNAI), vol. 5603, pp. 369–381. Springer, Heidelberg (2009)

Vetulani, Z.: Natural language based communication between human users and emergency center in critical situations. A short-text-message based decision assisting system POLINT-112-SMS. In: Vetulani, Z. (ed.) Proceedings of the 4th Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, 6–8 November 2009, Poznań, Poland, pp. 79–84. Wyd. Poznańskie, Poznań (2009)

Vetulani, Z., Obrębski, T.: Resources for extending the PolNet-Polish WordNet with a verbal component. In: Bhattacharyya, P., Fellbaum, C., Vossen, P. (eds.) Principles, Construction and Application of Multilingual Wordnets. Proceedings of the 5th Global Wordnet Conference, pp. 325–330. Narosa Publishing House, New Delhi (2010)

Vetulani, Z., Kubis, M., Obrębski, T.: PolNet – Polish WordNet: data and tools. In: Calzolari, N. (ed.) Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010), 19–21 May 2010, Valletta, Malta, (Proceedings), pp. 3793–3797. ELRA, Paris (2010a)

Vetulani, Z., Marcinak, J., Obrębski, J., Vetulani, G., Dabrowski, A., Kubis, M., Osiński, J., Walkowska, J., Kubacki, P., Witalewski, K.: Zasoby językowe i technologie przetwarzania tekstu. POLINT-112-SMS jako przykład aplikacji z zakresu bezpieczeństwa publicznego (in Polish) (Language resources and text processing technologies. POLINT-112-SMS as example of homeland security oriented application). Adam Mickiewicz University Press, Poznań (2010b). ISBN 978-83-232-2155-5

Kubis, M.: An access layer to PolNet – Polish WordNet. In: Vetulani, Z. (ed.) LTC 2009. LNCS, vol. 6562, pp. 444–455. Springer, Heidelberg (2011) (This is a revised version of the paper "An access layer to PolNet in POLINT-112-SMS" published in Proceedings of the 4th Language and Technology Conference, 6–8 November 2009, Poznan, Poland, pp. 437–441. Wydawnictwo Poznańskie, Poznań)

Vetulani, Z., Vetulani, G.: Through Wordnet to lexicon grammar (Abstract). In: Proceedings of the 5th Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, 25–27 November 2011, Poznań, Poland, p. 258. Wyd. Fundacja UAM, Poznań (2011)

Vetulani, Z., Marciniak, J.: Natural language based communication between human users and the emergency center: POLINT-112-SMS. In: Vetulani, Z. (ed.) LTC 2009. LNCS (LNAI), vol. 6562, pp. 303–314. Springer, Heidelberg (2011)

Vetulani, Z.: Wordnet based lexicon grammar for Polish. Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012), 23–25 May 2012, Istanbul, Turkey (Proceedings), pp. 1645–1649. ELRA, Paris (2012). ISBN 978-2-9517408-7-7. http://www.lrec-conf.org/proceedings/lrec2012/index.html

Vetulani, Z.: Language resources in a public security application with text understanding competence. A case study: POLINT-112-SMS. Proceedings of the LRPS Workshop at LREC 2012, 27 May 2012, Istanbul, Turkey. ELRA, Paris (2012). ISBN 978-2-9517408-7-7

Walkowska, J.: Modelowanie kompetencji dialogowej człowieka na potrzeby jej emulacji w zarządzających wiedzą systemach informatycznych współpracujących z wieloma użytkownikami (in Polish). Ph.D. thesis, IPI PAN, Luty 2012, Warszawa (2012)

Kubis, M.: A tool for transforming WordNet-like databases. In: Vetulani, Z., Mariani, J. (eds.) Human Language Technology Challenges for Computer Science and Linguistics, LTC 2011. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence), LNCS (LNAI), vol. 8387, pp. xx–yy. Springer, Heidelberg (2014)

Vetulani, Z., Kochanowski, B.: "PolNet - Polish WordNet" project: PolNet 2.0 - a short description of the release. In: Orav, H., Fellbaum, C., Vossen, P. (eds.) Proceedings of the Global Wordnet Conference, 2014, Tartu, Estonia, pp. 400–404. Ed. by Global Wordnet Association, Amsterdam (2014). ISBN 978-9-9493249-2-7

Vetulani, Z.: PolNet – Polish WordNet. In: Vetulani, Z., Mariani, J. (eds.) Human Language Technology Challenges for Computer Science and Linguistics, LTC 2011. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence), LNCS (LNAI), vol. 8387, pp. xx–yy. Springer, Heidelberg (2014)

Vetulani, Z., Vetulani, G.: Through Wordnet to lexicon grammar (Full text). In: Kakoyianni Doa, F. (ed.) Penser le lexique-grammaire: perspectives actuelles, pp. 531–545. Editions Honoré Champion, Paris (2014)

Other Publications

Vetulani, Z.: Linguistically motivated ontological systems. In: Callaos, N., Lesso, W., Schewe, K.-D., Atlam, E. (eds.) Proceedings of the 7th World Multiconference on Systemics, Cybernetics and Informatics, 27–30 July 2003, Orlando, FL, USA, vol. XII (Information Systems, Technologies and Applications: II), pp. 395–400. Int. Inst. of Informatics and Systemics (2003)

Gross, M.: Constructing lexicon-grammars. In: Sue Atkins, B.T., Zampolli, A. (eds.) Computational Approaches to the Lexicon, pp. 213–263. Oxford University Press, Oxford (1994)

Fillmore, C.J., Baker, C.F., Sato, H.: The FrameNet database and software tools. In: Proceedings of the Third International Conference on Language Resources and Evaluation, vol. IV. LREC, Las Palmas (2002)

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

Pease, A.: Ontology: A Practical Guide. Articulate Software Press, Angwin (2011)

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

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

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

Vetulani, G.: 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 (2012)

Vetulani, Z.: Towards a linguistically motivated ontology of motion: situation based synsets of motion verbs. In: Barr, V., Markov, Z. (eds.) Proceedings of the Seventheens International Florida Artificial Intelligence Research Society Conference (FLAIRS-04), pp. 813–817. AAAI Press, Menlo Park (2004)