plWordNet 2.3 emo and enWordNet 1.0: a demo

Marek Maziarz*, Maciej Piasecki*, Ewa Rudnicka*, Stan Szpakowicz†

* Department of Computational Intelligence, Wrocław University of Technology, Wrocław, Poland
mawroc@gmail.com, maciej.piasecki@pwr.wroc.pl, ewa.rudnicka78@gmail.com
† School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, Ontario, Canada
szpak@eecs.uottawa.ca

Abstract

We present plWordNet, the largest wordnet in existence today, and plWordNet emo, the largest Polish sentiment lexicon. We show examples of lexical units and semantic relations, the basic statistics of plWordNet, and the statistics and the evaluation of sentiment annotation. We also present an extended version of Princeton WordNet, along with a short description of the procedure of introducing new lemmas into it.

Keywords: wordnet, plWordNet, lexical units, sense relations, synonymy, synset

A wordnet is a network with *lexical units* (word senses) or *synsets* (sets of synonymous senses) as nodes, and lexico-semantic relations as edges[1]. Examples of relations are hypernymy (superclass-subclass), meronymy (part-whole) and antonymy (opposites). The incoming and outgoing relations define implicitly the meaning of a node; a wordnet also contains dictionary-style definitions and usage examples.

The work on WordNet (Fellbaum, 1998) began in the late 1980s at Princeton University. This first ever wordnet – a thesaurus, a dictionary organised by concepts, a taxonomy – has become a model for much vigorous development. In the past two and a half decades, hundreds of research teams have followed in the footsteps of WordNet's creators. Among them there is a team from Wrocław University of Technology with a resource under construction since 2006. Significantly, plWordNet is one of the few wordnets which are *not* the effect of translating WordNet. It has been built from the ground up, in a joint effort of lexicographers and computer scientists, to reflect the Polish lexical system.

The nodes in plWordNet are LUs, words together with their senses, variously interconnected by lexico-semantic relations from a well-defined relation set. For example, the synonymous lexical units *kot 2* and *kot domowy 1* 'cat, Felis domesticus' have the hypernym *kot 1* 'feline mammal, any member of the family Felidae' and such hyponyms as *dachowiec 1* 'alley cat' or *angora turecka 1* 'Turkish Angora'. An LU acquires its meaning from its relatedness to other LUs in the system; we can reason about it by considering relations in which it participates. Thus for example *kot 2* is defined as a kind of animal from the family Felidae, while *dachowiec 1* and *angora turecka 1* are kinds of Felis domesticus. LUs which enter the same lexico-semantic relations, but not the same derivational relations, are treated as synonyms and grouped into sets of units referred to as *synsets* in the wordnet parlance. For example, *kot 2* and *kot domowy 1* belong to the same synset (Figure 1).

In 2009, the first version of plWordNet with some 27,000 lexical units has been made freely available on the Internet. Today plWordNet describes almost 171,000 Polish nouns, verbs and adjectives, contains nearly 244,000 unique senses and 600,000 relation instances. It is not only the largest wordnet for Polish, but also already the largest wordnet in existence (see the statistics in Table 1).

wordnet	synsets	words (lemmas)	LUs	avs
GermaNet 10.0	101,371	119,231	131,814	1.30
WordNet 3.1	117,659	155,593	206,978	1.74
enWordNet 1.0	124,816	164,834	217,50	1.74
plWN 2.3	**184,240**	**170,834**	**244,286**	1.33

Table 1: The size of plWordNet 2.3 in synsets, lemmas and LUs, and average synset size (avs), compared to the very large WordNet 3.1 and GermaNet 10.0.

We now also present a pilot project to annotate plWordNet LUs manually with sentiment polarity, basic emotions and fundamental values. Basic emotions were adapted from the typology in (Plutchik, 1980), identified in his Wheel of Emotions: joy, trust, fear, surprise, sadness, disgust, anger, anticipation. Fundamental values, introduced into plWordNet from a model in (Puzynina, 1992), are utility – another's good – truth – knowledge – beauty – happiness, and their negative opposites.

We work with LUs, plWordNet's basic building blocks. So far, we have annotated about 30,000 nominal and adjectival LUs. The resulting lexicon is already one of the largest sentiment and emotion resources, in particular among those based on wordnets. We opted for manual annotation to ensure high accuracy, and to provide a reliable starting point for future semi-automated expansion (see Table 2).

Simultaneously, we worked hard on an extension of Princeton WordNet, called enWordNet. We decided to treat inter-lingual hyponymy (*I*-hyponymy) links between plWordNet and WordNet synsets as a starting (guiding) point for expansion. The lemmas of all plWordNet 'leaf' synsets linked by *I*-hyponymy relation to WordNet synsets

[1] The term *lexical unit* will be abbreviated to *LU* throughout this paper.

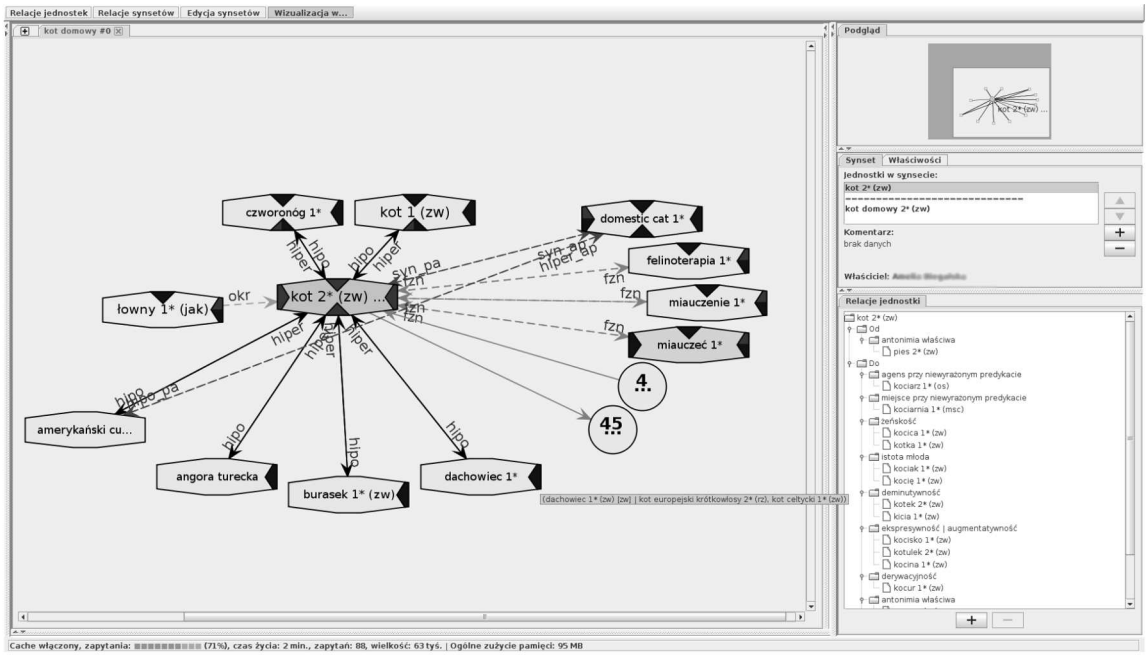

Figure 1: A screenshot from the WordnetLoom application with the synset {kot 2, kot domowy 1} in its centre.

ratio							
PoS	#	-s	-w	n	+w	+s	amb
N	19,625	11.29	8.78	69.06	3.24	2.88	4.74
Adj	11,573	9.89	11.22	58.85	9.21	5.60	5.24
Both	31,198	10.77	9.69	65.27	5.46	3.89	4.92
Fleiss' κ							
PoS	#	-s	-w	n	+w	+s	amb
N	19,625	0.961	0.915	0.976	0.864	0.930	0.868
Adj	11,573	0.958	0.935	0.960	0.919	0.976	0.935

Table 2: Experimental sentiment annotation of plWordNet 2.3 in percentage points ("ratio") and inter-annotator agreement, measured in Fleiss' κ ("Fleiss' κ"), for different types of sentiment polarity; -s, -w, n, +w, +s, amb (negative strong/weak, neutral, positive weak/strong, ambiguous). *N* stands for nouns, and *Adj* marks adjectives.

were filtered out (automatically translated) by a large cascade dictionary. The obtained list of translations was then filtered by the existing lemmas of WordNet. Next, the results of this filtering were divided into three groups: lemmas for we found equivalents whose lemmas were absent fromWordNet; lemmas for which we did not find any equivalents; lemmas for which we found equivalents whose lemmas were already present in WordNet. Lexicographers started with the first group, carefully verifying the suggestions against corpora and all available resources; then they moved to the second group, trying to find equivalents on their own (in all available resources); lastly, they investigated the third group, carefully verifying the existing mapping relations. Moreover, whenever lexicographers started work with a particular WordNet "nest", they were encouraged to look for its possible extensions on their own (do not limit themselves to cascade dictionary suggestions).

In numbers, the WordNet extension means adding 10,500 LUs, more than 7,000 lemmas and 9,000 synsets, which is 5% of WordNet's original size. However, WordNet contains not so many contemporary words (like *smartphone* or *tablet*, see Figure 2), so enWordNet succesfully fills the gap with contemporary vocabulary. This is very valuable in applications.

Wordnets are essential in natural language processing, for example in Information Retrieval, Question Answering and Sentiment Analysis. The larger a wordnet, the better. We compare plWordNet 3.0 (Maziarz et al., 2014), with Princeton WordNet 3.1 and, whenever possible, with another large wordnet, GermaNet (www.sfs.uni-tuebingen.de/GermaNet/). We show some standard graph statistics (Lewis, 2009): graph size (Table 1), average graph density (Table 3) and measure of hypernymy path for translation equivalents (Table 4) to find out whether plWordNet is of good quality.

The numbers say clearly that plWordNet is now the world's largest wordnet, comparable to the highly respected Princeton WordNet. It has denser relations, a stricter definition of synonymy (so, smaller synsets) and more linguistically-oriented hypernymy chains.

The continued growth of plWordNet and an extension of Princeton WordNethave been made possible by grants from the Polish Ministry of Science and Higher Educa-

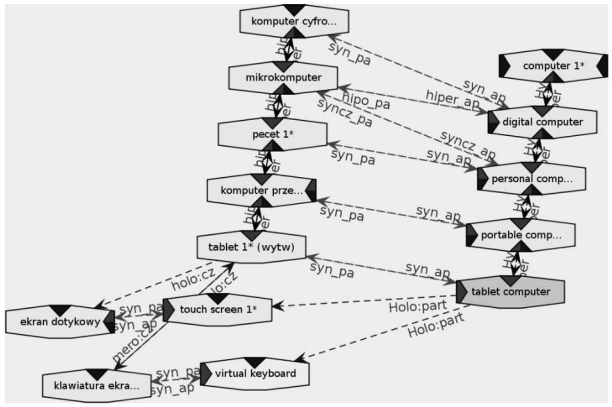

Figure 2: Screenshot from WordnetLoom application presenting the vicinity of synset {tablet computer 1, tablet 1} from enWordNet 1.0. One can see many newly introduced modern terms, *e.g., touchscreen* or *virtual keybord*, absent from the original WordNet 3.1.

POS	WordNet	plWordNet
nouns	2.5	3.17
verbs	3.32	3.95
adjectives	3.05	3.20
adverbs	0.88	—

Table 3: Synset relation density in WordNet 3.1 and in plWordNet 3.0 by part of speech (POS).

tion and from the European Union. We now work on it in the scope of the Clarin Poland project. We aim to build a conceptual dictionary fully representative of contemporary Polish, comparable with other very large wordnets. The extension of WordNet aims at increasing lexical coverage of the English wordnet. We have made an effort to ensure that plWordNetversion 3.0 and the very first version of WordNet have the same high quality as the best wordnets out there – Princeton WordNet, EuroWordNet (a joint initiative of a dozen or so members of the European Union) or GermaNet from Tübingen University. plWordNet and enWordNet are available free of charge for any applications, including commercial applications, on a licence modelled after that for Princeton WordNet.

Bibliography

Fellbaum, C. (ed.), 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Lewis, T.G., 2009. *Network Science: Theory and Applications*. Wiley.

Maziarz, M., M. Piasecki, E. Rudnicka, and S. Szpakowicz, 2014. plWordNet as the Cornerstone of a Toolkit of Lexico-semantic Resources. In *Proc. 7th International Global Wordnet Conference*.

Maziarz, Marek, Maciej Piasecki, Ewa Rudnicka, and Stanisław Szpakowicz, 2013a. Beyond the Transfer-and-Merge Wordnet Construction: plWordNet and Comparison with WordNet. In G. Angelova, K. Bontcheva, and R. Mitkov (eds.), *Proc. International Conference on Recent Advances in Natural Language Processing 2013*. Hissar, Bulgaria. ⟨http://lml.bas.bg/ranlp2013/docs/RANLP_main.pdf⟩.

path	avg.	std.
WordNet*up*	7.76	2.42
plWordNet *up*	5.71	3.33

Table 4: Average hypernym chain length (avg.) in WordNet 3.1 and in plWordNet 3.0 for Polish/English inter-lingual equivalents; std. denotes standard deviation.

FRC	≥1000	≥500	≥200	≥100	≥50
PWN	0.383	0.280	0.170	0.107	0.064
plWN	0.532	0.473	0.399	0.346	0.296

Table 5: Percentage of Princeton WordNet noun lemmas in *Wikipedia.en* and plWordNet lemmas in the plWordNet corpus. FRC is lemma frequency in the reference corpus.

Maziarz, Marek, Maciej Piasecki, and Stanisław Szpakowicz, 2012. Approaching plWordNet 2.0. In *Proc. 6th Global Wordnet Conference*. Matsue, Japan. ⟨http://nlp.pwr.wroc.pl/ltg/files/publications/paper%2042.pdf⟩.

Maziarz, Marek, Maciej Piasecki, and Stanisław Szpakowicz, 2013b. The chicken-and-egg problem in wordnet design: synonymy, synsets and constitutive relations. *Language Resources and Evaluation*, 47(3):769–796. ⟨link.springer.com/content/pdf/10.1007/s10579-012-9209-9.pdf⟩.

Piasecki, Maciej, Stanisław Szpakowicz, and Bartosz Broda, 2009. *A Wordnet from the Ground Up*. Wrocław University of Technology Press. ⟨www.eecs.uottawa.ca/~szpak/pub/A_Wordnet_from_the_Ground_Up.zip⟩.

Plutchik, Robert, 1980. *EMOTION: A Psychoevolutionary Synthesis*. Harper & Row.

Puzynina, Jadwiga, 1992. *Język wartości [The language of values]*. Scientific Publishers PWN.