plWordNet 3.0 – Almost There

Maciej Piasecki^A, Stan Szpakowicz^B, Marek Maziarz^A, Ewa Rudnicka^A
^{A} G4.19 Research Group, Department of Computational Intelligence
Wrocław University of Technology, Wrocław, Poland
^{B} School of Electrical Engineering and Computer Science
University of Ottawa, Ottawa, Ontario, Canada
^{A} maciej.piasecki@pwr.wroc.pl, mawroc@gmail.com,ewa.rudnicka78@gmail.com
^{B} szpak@eecs.uottawa.ca

Abstract

It took us nearly ten years to get from no wordnet for Polish to the largest wordnet ever built. We started small but quickly learned to dream big. Now we are about to release plWordNet 3.0-emo – complete with sentiment and emotions annotated – and a domestic version of Princeton WordNet, larger than WordNet 3.1 by nearly ten thousand newly added words. The paper retraces the road we travelled and talks a little about the future.

1 Wordnet makers' ambition

A respectable wordnet ought to be a fair model of the lexical-semantic system of the language it represents; a nearly comprehensive model is a dream worth pursuing. A wordnet linked to other wordnets, and to world knowledge, is a dream come true. This paper tells the story of plWordNet, a resource for Polish built over a decade of concentrated effort. Our wordnet is well published, but we are reaching a really large milestone, so we want take a bird's eye view of that decade.

We began cautiously. Our starting point in 2005 was a list of 10,000 most frequent lemmas in the IPI PAN Corpus of Polish, a mere quarter billion words from not quite balanced sources (Przepiórkowski, 2004). More than 30 person-years later, we are but a small step from completing the work on plWordNet 3.0-emo. With 177,003 lemmas, 255,733 lexical units, 193,286 synsets and more than 550,000 instances of relations, it is – in numbers – the largest wordnet created to date. Practically all its elements are in place, and the rollout is imminent. We think that it is an opportunity to present a synthetic picture of the whole endeavour.

The paper first recalls the initial fundamental assumptions, which have held astonishingly well, even if they had, inevitably, to be adjusted as our wordnet grew. We discuss the central lessons learned, and present the structure and statistics of plWordNet 3.0-emo. Finally, there is an overview of applications, and plans for the future.

2 Assumptions

We based the development of plWordNet on several unique assumptions, formulated *a priori*. They have been discussed at length in previous publications, notably (Piasecki et al., 2009; Maziarz et al., 2013c), so we will only recapitulate them briefly just to ease into the further discussion.

First and foremost, we believe that lexico-semantic systems of different languages differ in deep – and interesting – ways. That is why plWordNet, meant as *a precise description of the Polish lexical system*, had to be built in a way that avoided widespread influence of the material and structure of other wordnets. We were aware of the high cost of not simply translating Princeton WordNet, the only resource large enough for our ambition, but it felt most important to be faithful to the complex reality of our language.[1]

When the project began, there were no public-domain and no open-licence large electronic lexico-semantic resource for Polish.[2] We opted for a *corpus-based wordnet development process*. A very large corpus, the main knowledge source, is supplemented by a variety of linguistic substitution tests, mono-lingual dictionaries and other semantic language resources, encyclopædias, discussions among linguists, and the wordnet editors' linguistic and lexicographic intuition.

[1] In retrospect, this decision has been borne out by the scale of differences between plWordNet and WordNet when we got deep enough into the mapping between the two.

[2] There are scarcely any such resources even now (Vetulani et al., 2009; Miłkowski, 2007), unless one counts plWordNet ☺.

Corpus-based work, unaided by specialised software, would necessarily be rather slow. We assumed large-scale software support for *semi-automatic wordnet construction*, predicated on the availability of support tools for editing. Such tools were designed and built (and then perfected over the years) in parallel with fully manual construction of a small wordnet core to serve as the springboard for further expansion. This ensured much reduced workload for the editors and improved exploration of the corpus data. In many cases, the editor needs only to conform the support system's suggestions.[3]

It soon became clear that there were significant problems with making the usual synset definition operational, and with the consistency of the editors' decisions. We chose a smaller-grain basic element for plWordNet: the *lexical unit*.[4] A synset was then defined indirectly as a set of lexical units which share a number of *constitutive lexico-semantic relations and features* (Maziarz et al., 2013c). Relations between synsets are a notational abbreviation for the shared relations between lexical units grouped into those synsets. Constitutive relations, which define the structure of the wordnet, are complemented by relations which only link lexical units. Three categories of constitutive features are lexical registers, semantic classes of verbs and adjectives, and aspect. In this model, synonymy is also a derived concept: constitutive relation- and feature-sharing lexical units grouped into a synset are understood to be synonymous.

Finally, in the construction of plWordNet we tried to follow the principle of *a minimal commitment*, that is to say, to keep the number of assumptions small, to make plWordNet transparent to linguistics theories of meaning, and to shape it in a close relation to language data.

3 Lessons learned

3.1 Tools and organisation of work

Ten years of continuous wordnet development gave us a lot of practical experience which confirmed the initial assumptions.

The building of plWordNet was what can be termed a *corpus-based wordnet development process*. It starts with the lemmatisation of a large corpus and the extraction of the lemma frequency ranking. A top sublist of *new lemmas*, those not yet included in plWordNet, is selected for the given iteration of wordnet expansion. Typically, 6000-9000 new lemmas selected for an iteration meant 3-6 months of work. Each iteration processed lemmas in the same part of speech. We tried to "sanitise" every list by removing obvious non-words (mostly proper names), but serious cleaning would double the workload: it requires searching corpora and identifying potential senses.

Several tools examine the corpus to extract *knowledge sources* which help merge a new batch of lemmas with what is already in plWordNet: a Measure of Semantic Relatedness (MSR) and lists of lemma pairs potentially linked by hypernymy. The LexCSD system (Broda and Piasecki, 2011) extracts usage examples for the new lemmas. The extracted MSR was next used to cluster lemmas into semantically motivated groups we call *packages*, each package assigned to one editor. A package is clearly homogenous; usually, 2-3 domains are most prominent (lemmas were grouped by dominating senses), so the editor can stay focused. The acquired knowledge sources were input to the WordnetWeaver system (Piasecki et al., 2009) which, for each new lemma, automatically suggests the number and location in the network of lexical units. The suggestions are visually presented in the wordnet editing system WordnetLoom (Piasecki et al., 2010).

The plWordNet team consists of rank-and-file editors and coordinators.[5] Before tackling lemmas in any of four parts of speech, we prepared guidelines with detailed relation definitions and substitution tests. A coordinator entered the definitions and tests into WordnetLoom, and trained the editors. The coordinator assigns lemmas to editors in batches, performs selective verification, answers questions, refines the guidelines, and monitors the pace and progress of the editors' work.

For frequent lemmas, the editor uses supporting tools in a specified order of importance: WordnetWeaver suggestions; corpus browsers; usage

[3] Software support has also greatly assisted in the mapping between plWordNet and Princeton WordNet. Likewise, a mapping to knowledge resources, notably to ontologies, had to be built semi-automatically from scratch.

[4] A lexical unit is understood here as a triple: (lemma, part of speech, sense identifier). A lemma is the basic morphological form of a word. Each lexical unit represents a unique word sense.

[5] At the height of plWordNet development, several coordinators supervised a small group of editors each. Separate teams work on plWordNet-to-WordNet mapping, and on sentiment annotation. All this allows cross-checking: the teams exchange information about likely errors.

examples generated automatically by LexCSD and the induced senses they represent; lists of highly related lemmas according to MSR; existing electronic dictionaries, lexicons, encyclopaedias; and, last but not least, the linguistic intuition of the editor and the team. The importance of WordnetWeaver and MSR dropped for lower-frequency lemmas. In the case of nouns editors tend to use dictionaries as the main source, but still remember the other sources. Adjectives and adverbs are much less richly described in the existing dictionaries, so LexCSD examples and corpus browsers became the primary tools.

Before adding any relation instance to the wordnet, WordnetLoom presents the appropriate substitution test with the variable slots filled by the lexical units of the two synsets. The instantiated substitution test reminds about the constraints included in the relation definition, likely improving the consistency of the editors' definitions. Similarly, consistency increases with the use of the same supporting tools in the same order.

3.2 The role of corpora

Corpus-based development is surely slower and more costly than the merge method based on the previously existing lexical resources, but it is the only method which allows going beyond the existing dictionaries, often closely related. Corpus-based development also promotes a wordnet's better coverage of lemmas described and lexical units, assuming that the procedure recapped above is carefully followed. Obviously, a lot depends on the type of corpus. We aimed at building a comprehensive wordnet, so we tried to acquire or collect as large a corpus as possible. We made a practical assumption that the larger the corpus and the more diverse its text sources, the more balanced and representative the corpus becomes.

The development of plWordNet 1.0 relied on the IPI PAN Corpus (IPIC) (Przepiórkowski, 2004), ca. 260 million tokens, the first publicly available large corpus of Polish.[6] IPIC represents a range of genres, biased towards parliamentary documents and scientific literature. That is why we put much effort into collecting corpora and texts, and combining them with IPIC.

The work on plWordNet 2.1 built upon a plWordNet corpus of 1.8 billion tokens, recently expanded to almost 4 billion tokens. This merged corpus encompasses IPIC, the corpus of text from the newspaper *Rzeczpospolita* (Weiss, 2008) and Polish Wikipedia; it is complemented by texts collected from the Internet, filtered according to the percentage of unrecognised words by Morfeusz (Woliński, 2006), with duplicates removed with respect to the whole corpus.

Finally, plWordNet 3.0 describes all lemmas with 30+ occurrences in 1.8 billion words, as well as a significant number of those less frequent.[7] At the final stage of work on plWordNet 3.0, we plan to add missing lemmas with the frequency 30+ from the 4-billion-token corpus.

3.3 The underlying model

The strategy of making the lexical unit the basic building block helped us formulate definitions of relations, and substitution tests for those relations, so they refer primarily to language data and the distribution of lemmas in use examples. We could also refer to the linguistic tradition in defining lexico-semantic relations better matching the background of our editors. We are convinced that the use of elaborate relation definitions, substitution tests and the procedure of lexicographic work have improved the mutual understanding of the plWordNet model among the members of the linguistic team, as well as the consistency of editing decisions across the pool of editors.

The model of plWordNet, based on the sharing of constitutive relations and features, allowed us to write up and implement an operational definition of the synset. Still, specific leaves deep in the wordnet hypernymy tree often could not be easily separated into different synsets without referring to some notion of synonymy (or – more important in practice – to the absence of synonymy). We "pinned it down" as a combination of two parallel hyponymy relations. We think that the need for synonymy in wordnet editors' everyday work can be reduced in the future as the list of relations grows. That was what happened with verbs, adjectives and adverbs, for which we introduced, *e.g.,* several cross-categorial constitutive relations.

3.4 The progress of work

We deliberately avoided putting non-lexical elements in plWordNet, a lexical resource *par ex-*

[6] Oddly, it is even now the only freely available corpus of Polish. It is a pity that the newer and larger National Corpus of Polish (Przepiórkowski et al., 2012) is not all in the public domain (http://nkjp.pl/).

[7] Editors were free to add any existing lemma, after checking corpora (Przepiórkowski et al., 2012) and the Internet.

cellence. For example, we only included proper names from which frequent lexical units are derived; other proper names are kept in a separate large lexicon mapped onto plWordNet. We have also developed an elaborate procedure for assessing the lexicality of multiword expressions. We made an exception for "artificial" (non-lexical) synsets first proposed for GermaNet (Hamp and Feldweg, 1997). They usually make a wordnet's hypernymy structure more readable for humans. The added artificial nodes also help editors maintain the hypernymy structure. Consequently, a significant number of *artificial lexical units* (language expressions) have been placed in singleton synsets. Such synsets and lexical units, clearly marked, can be removed or made transparent, if needed. They are not treated as part of the lexical system described by the wordnet.

The WordnetWeaver system implements a complex frequency-based method of wordnet expansion[8] (Piasecki et al., 2013). The method worked fine in the first phase of plWordNet development, for frequent lemmas, mostly nouns. With the move to less frequent lemmas, the importance of WordnetWeaver waned. Its Measure of Semantic Relatedness (MSR), an essential knowledge source, proves useful for lemmas occurring 200+ times (an observed empirical rule); below 100 occurrences, it begins to produce many accidental associations. The thresholds are even higher for verbs, if the description of their occurrences is not based on the output of a reliable parser.

While we abandoned WordnetWeaver for less frequent lemmas, several of its components remain in use. Most important, even if the MSR's quality decreases, it helps automated semantic clustering of lemmas in aid of assigning work to individual editors. Semantically motivated packages for this purpose, even if imperfect, handily beat such schemas as alphabetic order. Also, the LexCSD system automatically extracts use examples meant to represent various senses of a new lemma. LexCSD clusters all occurrences of the lemma, and tries first to identify occurrence groups representing different senses, and then to find the most prominent example in each group.

Examples extracted by LexCSD are also presented in WordnetLoom. Such examples have become the first knowledge source which plWordNet editors consult when they work on adjectives and adverbs. Existing Polish dictionaries neglect both categories, so we rely on corpus-derived examples. Lexico-syntactic patterns used for the extraction of lemma pairs potentially linked by a given relation also apply to less frequent words; the practice shows, however, that they are also less frequent in language expressions matching the patterns. Automated methods were very helpful in expanding derivational relations in plWordNet (Piasecki et al., 2012a; Piasecki et al., 2012b). Regardless of which automatic method was used, the results were always verified by human editors and revised if necessary.

The manual mapping of plWordNet onto Princeton WordNet has incurred a high labour cost, even though we deliberately stayed away – for now – from the opposite direction (Rudnicka et al., 2012). We built an automated system to suggest inter-lingual links (Kędzia et al., 2013). Its precision is acceptable, but too low to let the results stand without intervention. We have also introduced several inter-lingual relations (Rudnicka et al., 2012) in order to cope with nontrivial differences between the two wordnets. All that investment was worth the price. The bilingual resource we now have is unique in scale (two largest wordnets, over 150,000 interlingual links between synsets) and nature (two wordnets based on slightly different models). The mapping opens many interesting paths for further exploration.

Early on, we assumed tacitly that glosses were not part of the relational model of language which our wordnet represented. We still think that it is better first to invest in building a larger gloss-free wordnet than to construct a much smaller but more lexicographically complete resource.[9] A wordnet describes the meaning of a lexical unit *via* its network of lexico-semantic relations. Inevitably, though, as plWordNet gained popularity (through its Web page and mobile application), we soon noted that glosses help non-specialist users understand the meaning of wordnet entries. It is a technicality, perhaps, but glosses also help wordnet editors see clearly the editing decisions made by other members of the team: glosses serve as a form of control information. Similarly, use examples help, and appear more important for Natural Language Engineering applications of plWordNet.

[8] automated, but subject to editors' final approval

[9] Come to think of it, glosses in Princeton WordNet were an afterthought, too. ☺

4 The structure of plWordNet 3.0-emo

Maziarz et al. (2013a) presented plWordNet 2.1. In most ways, plWordNet 3.0 is just better and larger, as planned two years ago (Maziarz et al., 2014). In comparison to version 2.1:

- noun and adjective sub-databases have grown very substantially – see the statistics in Section 5; the verb, already a large list, have been only amended;
- the set of adjective relations has been revised, while only minor changes were introduced for nouns and verbs;
- a new adverb sub-database has been constructed from scratch with the help of a semi-automatic method based on exploring derivational relations and mapping between adjective relations and adverb relations;
- an elaborate procedural definition of Multi-word Lexical Units was designed (Maziarz et al., 2015), together with a work procedure supported by the semi-automatic system for collocation extraction and their further editing as potential candidates;
- the plWordNet-to-WordNet mapping has been very significantly expanded to adjectives, with coverage vastly increased to 151,200 interlingual links of various types (38,471 I-synonymy links);
- the constructed bilingual mapping was used to build a rule-based automated procedure of mapping plWordNet to SUMO (Pease, 2011; Kędzia and Piasecki, 2014).

4.1 Mapping to WordNet

To this planned development, we added two derived resources. While mapping onto Princeton WordNet, we observed that the most frequent inter-lingual relation is I-hyponymy (over twice more frequent than I-synonymy). That is to say, there were no counterparts in WordNet 3.1 for many specific lexical units in plWordNet. The cause: differences in coverage between both wordnets rather than any major differences in lexicalisation between Polish and English (Maziarz et al., 2013a), even though we dutifully checked English dictionaries and corpora for direct translations. Now, I-hyponymy is more vague – gives us less useful information for language processing – than I-synonymy. That is why we decided to add material to WordNet 3.1. The result is a resource we call enWordNet 0.1, included in the plWordNet distribution as a large bilingual system. It has been built by adding to WordNet 3.1 about 8,000 new noun lemmas (9,000 noun lexical units).[10]

We aimed to improve the mapping of plWordNet (by adding to WordNet the missing corresponding entries), and then to replace I-hyponymy with I-synonymy as much as possible. This could be done simply by translating plWordNet synsets into English and putting the translations in enWordNet,[11] but we resisted that temptation.

We decided to let I-hyponymy guide expansion. The lemmas of all plWordNet 'leaf' synsets linked by I-hyponymy to WordNet synsets were automatically translated by a large cascade dictionary. The translations were then filtered by the existing WordNet lemmas and divided into three groups, lemmas for which the dictionary found: (i) equivalents whose lemmas were absent from WordNet; (ii) no equivalents; (iii) equivalents whose lemmas were already present in WordNet. Editors started with the first group, carefully verifying the suggestions with corpora, especially BNC (BNC, 2007) and ukWaC (Ferraresi et al., 2008), and all available resources. For the second group, they tried to find equivalents on their own (in all available resources). Finally, they investigated the third group, checking the existing mapping relations. Whenever editors started work with a particular WordNet 'nest', they were encouraged to look for its possible extensions on their own, not just limit themselves to the cascade dictionary suggestions.

We began with nouns. That segment of Princeton WordNet figures in applications more often than other parts of speech. Also, our experience with developing plWordNet suggested that adding to the nouns in WordNet would be relatively easy. We used the same set of relations as in Princeton WordNet but, following the plWordNet practice, the relations have been specified by definitions and substitution tests in the WordnetLoom editing system. The editor team consisted of graduates of English philology and native speakers.

In the first phase, we used bilingual dictionaries to select from the list those lemmas which appeared to be missing translation equivalents for plWordNet synsets lacking I-synonymy. Even so, the processing of the selected lemmas was in-

[10] The estimated target size is 10,000 new nouns.

[11] That would mean applying the transfer method in an "unorthodox" direction. One normally translates English synsets into whatever language one is building a wordnet for.

dependent of their potential Polish counterparts. Only after new lexical units had been added to enWordNet would the interlingual mapping be modified or expanded. For each English lemma, the editors identified its senses by searching for use examples in the corpora. We allowed into enWordNet only lexical units with 5+ occurrences, supported by examples.

In the second phase, we used the rest of the lemma list extracted from the corpora going through the lemmas of decreasing frequency.

4.2 Sentiment and emotions

Section 6 shows how plWordNet has become an important resource for language engineering applications in Polish. A notable exception were applications in sentiment analysis, despite their growing importance among research and commercial systems. That is why we decided to annotate manually a substantial part of plWordNet with sentiment polarity, basic emotions and fundamental values (Zaśko-Zielińska et al., 2015). The suffix "-emo" in the name of this plWordNet version signals the presence of this annotation. All in all, 19,625 noun lexical units and 11,573 adjective lexical units received two manual annotations. The team consisted of linguists and psychologists, whose coordinator was tasked with breaking ties. Each lexical unit was annotated with:

- its sentiment polarity (positive, negative, ambiguous) and its intensity (strong, weak);
- basic emotions associated with it: joy, trust, fear, surprise, sadness, disgust, anger, anticipation (Plutchik, 1980);
- fundamental human values associated with it: *użyteczność* 'utility', *dobro drugiego człowieka* 'another's good', *prawda* 'truth', *wiedza* 'knowledge', *piękno* 'beauty', *szczęście* 'happiness' (all of them positive), *nieużyteczność* 'futility', *krzywda* 'harm', *niewiedza* 'ignorance', *błąd* 'error', *brzydota* 'ugliness', *nieszczęście* 'misfortune' (all negative) (Puzynina, 1992).

The annotation of nouns encompassed those hypernymy sub-hierarchies which we expected to include lexical units with non-neutral sentiment polarity. Those were the sub-hierarchies for affect, feelings and emotions, nouns describing people, features of people and animals, artificial lexical unit *events rated negatively*, *evaluated as negative* and the sub-hierarchy of entertainment. The adjectival part of plWordNet was in major expansion during that time, so we only annotated the parts for which the expansion had been completed.

POS	**synsets**	**lemmas**	**LUs**	**avs**
N-PWN	82,115	117,798	146,347	1.78
N-enWN	88,381	125,819	155,437	1.76
N-plWN	123,985	126,746	167,243	1.35
V-PWN	13,767	11,529	25,047	1.81
V-enWN	13,789	11,540	25,061	1.82
V-plWN	21,669	17,398	31,841	1.47
A-PWN	18,156	21,785	30,004	1.65
A-enWN	18,185	21,808	30,072	1.65
A-plWN	39,204	27,041	45,899	1.17
Adv-PWN	3,625	4,475	5,592	1.54
Adv-enWN*	3,625	4,475	5,592	1.54
Adv-plWN	8,080	5,719	10,416	1.29
GermaNet	101,371	119,231	131,814	–
PWN	117,659	155,593	206,978	1.74
enWN	124,266	164,032	216,623	1.73
plWN	**193,286**	**177,003**	**255,733**	1.32

Table 1: The count by part of speech (PoS) of Noun/Verb/Adjective synsets, lemmas and lexical units (LUs), and average synset size (avs), in PWN 3.1 (PWN), enWordNet 0.1 (enWN), plWordNet 3.0 (plWN) and GermaNet 10.0 (www.sfs.uni-tuebingen.de/GermaNet/).

*This part of WordNet remains to be extended.

It is worth emphasizing that the amount of manual annotation is several times higher than in other wordnets annotated with sentiment. This pilot study can be a good starting point for semi-automated annotation of the whole plWordNet.

5 Statistics

Wordnets are treated as basic lexical resources, so their sizes matter a lot for potential applications. See Table 1 for the general statistics in plWordNet 3.0-beta-emo and a comparison with the other very large wordnets. We note that plWordNet has been consistently expanded in all parts of speech (PoS). The ratio between the size of plWordNet and Princeton WordNet is roughly the same for all PoS. The development of enWordNet has been intentionally concentrated on nouns.

Moreover, plWordNet has become larger than all modern dictionaries of general Polish in terms of the entries included: 130k (Zgółkowa, 1994 2005), 125k [180k lexical units] (Doroszewski, 1963 1969), 100k [150k lexical units] (Dubisz, 2004), 45k [100k lexical units] (Bańko, 2000). One of the main reasons is that those dictionaries do not contain many specialised words and senses from science, technology, culture and so on. Such material, however, is appropriate for a wordnet due to its applications in processing of texts of many genres coming from different sources, including the Internet. We could also observe that lemma lists added to plWordNet (based on the corpus) included quite a few words that are now frequent, but not described in those dictionaries.

The largest ever Polish dictionary, from the early 1900s, has 280k entries (Karłowicz et al., 1900 1927; Piotrowski, 2003, p. 604) and is still much larger than plWordNet, but it contains many archaic words, perhaps useful in the processing of texts from specialised domains. The achieved size of plWordNet has already exceeded the target size estimated for it considering a corpus of 1.8 billion words (Maziarz et al., 2014).

Lexico-semantic relations are the primary means of description of lexical meanings represented in a wordnet by synsets. The average number of relation links per synset, which is called *relation density*, tells us about the average amount of information provided by the wordnet for a single lexical meaning. Table 2 compares the relation density in Princeton WordNet and plWordNet for different parts of speech (obligatory inverse relations have been excluded from the count).[12] The relation density is higher in plWordNet for all parts of speech. We can name two reasons for this difference: smaller synsets in plWordNet on average, see Table 1, and the assumed way of defining synsets by the constitutive relations – more relations are needed to distinguish different synsets (*i.e.,* lexical meanings). However, plWordNet has a rich set of relations (more than 40 main types and 90 sub-types). Some of them have originated from the derivational relations. That can also increase the relation density.

If a wordnet is treated as a reference source, we expect to find in it most of the lemmas from the processed text. The complete coverage is not possible, but the higher it is, the more information a wordnet provides for the analysed text. Table 3 compares the coverage of Princeton WordNet and plWordNet for two corpora of a comparable size. From both corpora, two lemma frequency lists were extracted. Both corpora were first morphosyntactically tagged and only lemmas of the parts of speech described in wordnets were taken into account. For Polish, we worked with the plWordNet corpus (version 7) of ≈1.8 billion words from several available corpora (see section 3.2), supplemented by texts collected from the Internet. As an English corpus, we took texts from the English Wikipedia, ≈1.2 billion words, a size similar to that of the plWordNet corpus.[13]

POS	**Princeton WordNet**	**plWordNet**
nouns	2.5	3.17
verbs	3.32	3.95
adjectives	3.05	3.20
adverb	0.88	4.53

Table 2: Synset relation density in Princeton WordNet 3.1 and in plWordNet 2.0 by part of speech.

FRC	≥1000	≥500	≥200	≥100	≥50
PWN	0.383	0.280	0.170	0.107	0.064
plWN	0.732	0.644	0.515	0.416	0.327

Table 3: Percentage of Princeton WordNet noun lemmas in *Wikipedia.en* and plWordNet (plWN) lemmas in the plWordNet corpus. FRC is lemma frequency in the reference corpus.

The coverage is much higher for plWordNet, but the corpora differ. Many more specialised and rare words appear in English Wikipedia than in the Polish corpus. Even so, the statistics bode well for plWordNet's potential in applications. The coverage for the most frequent words ($\geq$ 1000) is not 100% because the list includes many proper names and misspelled words recognised by the tagger as common words. In comparison with plWordNet 2.1 (Maziarz et al., 2013b), the coverage of less frequent words increased significantly, because the development of plWordNet moves towards the bottom of the frequency ranking list.

The average polysemy – the ratio of lexical units to lemmas – is higher in plWordNet than in WordNet both for nouns (1.32 *vs* 1.24) and adjectives (1.71 *vs* 1.38). The difference is lower than in

[12] The relation structures differ among the parts of speech, so we do not show relation density for the whole wordnets.

[13] We used the plWordNet corpus to build the wordnet and to evaluate it. This may suggest a biased comparison. WordNet is evaluated on a corpus unrelated to its development, so only a qualitative comparison is warranted. Regardless, both wordnets more willingly absorb frequent than infrequent lemmas (Maziarz et al., 2013b).

plWordNet 2.1: we added more specific monosemous lemmas as a result of the focus given to lexical units and the tendency to describe exhaustively all existing lexical units for a given lemma. For verbs we have 1.83 *vs* 2.17, maybe because of aspect and rich derivation in Polish verbs.

The comparison of hypernymy path lengths did not change much from plWordNet 2.1 (Maziarz et al., 2013a). WordNet's much longer paths are caused by the elaborate topmost part of its hypernymy hierarchy; plWordNet has ≈100 linguistically motivated hypernymy roots.[14]

6 Applications

Wordnet-building costs a lot of public money, so as a rule the effect should be free for the public use. This good rule, grounded in Princeton WordNet's practice, is central for languages other than English, still less resourced. The availability of plWordNet on the WordNet-style open licence has stimulated, over the years, many interesting applications in linguistic research, language resources and tools, scientific applications, commercial applications and education.

The plWordNet Web page and Web service have had tens of thousands of visitors, and hundreds of thousands of searches. There are over 100 citations and over 700 users, individual and institutional, who optionally registered when downloading the plWordNet source files. Most of the registered users described the intended use of plWordNet, and a rich tapestry it is. The limited space only allows us to single out a handful in citations.

First of all, plWordNet has been applied in linguistic research: valency frame description and automated verb classification; verb analysis for semantic annotation in a corpus of referential gestures; contrastive/comparative studies, *etc*.

Increasingly often, plWordNet is treated as a large monolingual and bilingual dictionary, *e.g.,* in text verification during editing or as a source of meta-data for publications. Miłkowski (2010) included plWordNet among the dictionaries in a proofreading tool and as a knowledge source for an open Polish-English dictionary, which many translators and translation companies say they use. Open Multilingual Wordnet (Bond, 2013) now includes plWordNet. It is referred to in several other projects on wordnets and semantic lexicons (Pedersen et al., 2009; Lindén and Carlson, 2010; Borin and Forsberg, 2010; Mititelu, 2012; Zafar et al., 2012; Šojat et al., 2012). Practical machine translation systems use plWordNet. We are aware of applications in measuring translation quality and building the MT component embedded in an application supporting English teaching to children.

There are more research and commercial projects, both under way and announced by plWordNet users. They include ontology building and linking, information retrieval, question answering, text mining, semantic analysis, terminology extraction, word sense disambiguation (WSD), text classification, sentiment analysis and opinion mining, automatic text summarisation, speech recognition, or even the practice of aphasia treatment.

7 The lexicographer's work is never done

When in 2012 we established the target size of plWordNet 3.0, we were convinced that we would go to limits of the Polish lexical system. We now see that – even if major paths have been explored – we are discovering numerous smaller paths going deeper into the system.

The Polish side of plWordNet could have more relation links per synset. The constitutive relations do not differentiate all hypernymy leaves yet. There are cross-categorial relations, more numerous than in many other wordnets, but still not enough for WSD or semantic analysis. The connection to the valency lexicon could be tighter. The description of verb derivation (as highly productive in Polish as in other Slavic languages) needs much more work, and so do some relations, *e.g.,* meronymy. More information useful for WSD could be introduced, *e.g.,* further glosses or links to external sources like Wikipedia. Finally, for applications in translation (manual and machine-based) we must not only complete the mapping to WordNet, but also go inside synsets, *i.e.,* map lexical units. We are fortunate to have so much more intriguing work to do.

Acknowledgment

Work supported by the Polish Ministry of Education and Science, Project CLARIN-PL, the European Innovative Economy Programme project POIG.01.01.02-14-013/09, and by the EU's 7FP under grant agreement No. 316097 [ENGINE].

[14] They do not have hypernyms according to the definitions assumed in plWordNet.

References

[Bańko2000] Mirosław Bańko, editor. 2000. *Inny słownik języka polskiego PWN [Another dictionary of Polish]*, volume 1-2. Polish Scientific Publishers PWN.

[BNC2007] BNC. 2007. The British National Corpus, version 3 (BNC XML Edition). Distributed by Oxford University Computing Services on behalf of the BNC Consortium.

[Bond2013] Francis Bond. 2013. Open Multilingual Wordnet. http://compling.hss.ntu.edu.sg/omw/.

[Borin and Forsberg2010] Lars Borin and Markus Forsberg. 2010. From the People's Synonym Dictionary to fuzzy synsets – first step. In *Proc. LREC 2010*.

[Broda and Piasecki2011] Bartosz Broda and Maciej Piasecki. 2011. Evaluating LexCSD in a large scale experiment. *Control and Cybernetics*, 40(2):419–436.

[Calzolari et al.2012] Nicoletta Calzolari et al., editor. 2012. *Proc. Eighth International Conference on Language Resources and Evaluation (LREC-2012)*. European Language Resources Association.

[Doroszewski1963 1969] Witold Doroszewski, editor. 1963–1969. *Słownik języka polskiego [A dictionary of the Polish language]*. Państwowe Wydawnictwo Naukowe.

[Dubisz2004] Stanisław Dubisz, editor. 2004. *Uniwersalny słownik języka polskiego [A universal dictionary of Polish], electronic version 1.0*. Polish Scientific Publishers PWN.

[Ferraresi et al.2008] Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proc. 4th Web as Corpus Workshop (WAC-4)*, pages 47–54.

[Hamp and Feldweg1997] Birgit Hamp and Helmut Feldweg. 1997. GermaNet – a Lexical-Semantic Net for German. In *Proc. ACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15. Madrid.

[Isahara and Kanzaki2012] Hitoshi Isahara and Kyoko Kanzaki, editors. 2012. *Advances in Natural Language Processing: Proc. 8th International Conference on NLP, JapTAL*, volume 7614 of *Lecture Notes in Artificial Intelligence*. Springer-Verlag.

[Karłowicz et al.1900 1927] Jan Karłowicz, Adam Antoni Kryński, and Władysław Niedźwiedzki, editors. 1900-1927. *Słownik języka polskiego [A dictionary of the Polish language]*. Nakładem prenumeratorów i Kasy im. Józefa Mianowskiego [Funded by subscribers and Józef Mianowski Fund], Warsaw.

[Kędzia et al.2013] Paweł Kędzia, Maciej Piasecki, Ewa Rudnicka, and Konrad Przybycień. 2013. Automatic Prompt System in the Process of Mapping plWordNet on Princeton WordNet. *Cognitive Studies*. to appear.

[Kędzia and Piasecki2014] Paweł Kędzia and Maciej Piasecki. 2014. Ruled-based, interlingual motivated mapping of plwordnet onto sumo ontology. In Nicoletta Calzolari et al., editor, *Proc. Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association.

[Lindén and Carlson2010] Krister Lindén and Lauri Carlson. 2010. FinnWordNet – WordNet på finska via översättning. *LexicoNorcdica*, 17.

[Maziarz et al.2013a] Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, and Stan Szpakowicz. 2013a. Beyond the Transfer-and-Merge Wordnet Construction: plWordNet and a Comparison with WordNet. In G. Angelova, K. Bontcheva, and R. Mitkov, editors, *Proc. International Conference on Recent Advances in Natural Language Processing*. Incoma Ltd.

[Maziarz et al.2013b] Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, and Stan Szpakowicz. 2013b. Beyond the transfer-and-merge wordnet construction: plWordNet and a comparison with WordNet. In *Proc. International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 443–452. INCOMA Ltd. Shoumen, BULGARIA.

[Maziarz et al.2013c] Marek Maziarz, Maciej Piasecki, and Stanisław Szpakowicz. 2013c. The chicken-and-egg problem in wordnet design: synonymy, synsets and constitutive relations. *Language Resources and Evaluation*, 47(3):769–796.

[Maziarz et al.2014] Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, and Stan Szpakowicz. 2014. plWordNet as the Cornerstone of a Toolkit of Lexico-semantic Resources. In *Proc. Seventh Global Wordnet Conference*, pages 304–312.

[Maziarz et al.2015] Marek Maziarz, Stan Szpakowicz, and Maciej Piasecki. 2015. A procedural definition of multi-word lexical units. In *Proc. RANLP 2015*, page to appear.

[Miłkowski2010] Marcin Miłkowski. 2010. Developing an open-source, rule-based proofreading tool. *Software – Practice and Experience*.

[Mititelu2012] Verginica Barbu Mititelu. 2012. Adding Morpho-semantic Relations to the Romanian Wordnet. In *Proc. LREC 2012*.

[Miłkowski2007] Marcin Miłkowski. 2007. Open Thesaurus - polski Thesaurus. http://www.synomix.pl/.

[Pease2011] Adam Pease. 2011. *Ontology - A Practical Guide*. Articulate Software Press.

[Pedersen et al.2009] Bolette Sandford Pedersen, Sanni Nimb, Jørg Asmussen, Nicolai Hartvig Sørensen, Lars Trap-Jensen, and Henrik Lorentzen. 2009. DanNet: the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation*.

[Piasecki et al.2009] Maciej Piasecki, Stanisław Szpakowicz, and Bartosz Broda. 2009. *A Wordnet from the Ground Up*. Wrocław University of Technology Press. http://www.eecs.uottawa.ca/ szpak/pub/A_Wordnet_from_the_Ground_Up.zip.

[Piasecki et al.2010] Maciej Piasecki, Michał Marcińczuk, Adam Musiał, Radosław Ramocki, and Marek Maziarz. 2010. WordnetLoom: a Graph-based Visual Wordnet Development Framework. In *Proc. Int. Multiconf. on Computer Science and Information Technology – IMCSIT 2010, Wisła, Poland, October 2010*, pages 469–476.

[Piasecki et al.2012a] Maciej Piasecki, Radosław Ramocki, and Marek Maziarz. 2012a. Automated Generation of Derivative Relations in the Wordnet Expansion Perspective. In *Proc. 6th Global Wordnet Conference*.

[Piasecki et al.2012b] Maciej Piasecki, Radosław Ramocki, and Paweł Minda. 2012b. Corpus-based semantic filtering in discovering derivational relations. In Allan Ramsay and Gennady Agre, editors, *Proc. 15th International Conference on Artificial Intelligence: Methodology, Systems, Applications*, volume 7557 of *Lecture Notes in Computer Science*, pages 14–42. Springer.

[Piasecki et al.2013] Maciej Piasecki, Radosław Ramocki, and Michal Kaliński. 2013. Information spreading in expanding wordnet hypernymy structure. In *Proc. International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 553–561. INCOMA Ltd. Shoumen, BULGARIA.

[Piotrowski2003] Tadeusz Piotrowski, 2003. *Współczesny język polski [Contemporary Polish], edited by Jerzy Bartmiński*, chapter Słowniki języka polskiego [Dictionaries of Polish]. Marie Curie-Sklodowska University Press.

[Plutchik1980] Robert Plutchik. 1980. *EMOTION: A Psychoevolutionary Synthesis*. Harper & Row.

[Przepiórkowski et al.2012] Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors. 2012. *Narodowy Korpus Języka Polskiego [in Polish]*. Wydawnictwo Naukowe PWN. http://nkjp.pl/settings/papers/NKJP_ksiazka.pdf.

[Przepiórkowski2004] Adam Przepiórkowski. 2004. *The IPI PAN Corpus: Preliminary version*. Institute of Computer Science, Polish Academy of Sciences.

[Puzynina1992] Jadwiga Puzynina. 1992. *Język wartości [The language of values]*. Scientific Publishers PWN.

[Rudnicka et al.2012] Ewa Rudnicka, Marek Maziarz, Maciej Piasecki, and Stan Szpakowicz. 2012. A Strategy of Mapping Polish WordNet onto Princeton WordNet. In *Proc. COLING 2012, posters*, pages 1039–1048.

[Šojat et al.2012] Krešimir Šojat, Matea Srebačić, and Marko Tadić. 2012. Derivational and Semantic Relations of Croatian Verbs. *Journal of Language Modelling*, 0(1):111–142.

[Vetulani et al.2009] Zygmunt Vetulani, Justyna Walkowska, Tomasz Obrębski, Jacek Marciniak, Paweł Konieczka, and Przemysław Rzepecki. 2009. An Algorithm for Building Lexical Semantic Network and Its Application to PolNet – Polish WordNet Project. In Zygmunt Vetulani and Hans Uszkoreit, editors, *Human Language Technology. Challenges of the Information Society, Third Language and Technology Conf., Poznań, Revised Selected Papers*, LNCS 5603, pages 369–381. Springer.

[Weiss2008] Dawid Weiss. 2008. Korpus Rzeczpospolitej [Corpus of text from the online edition of "Rzeczpospolita"]. http://www.cs.put.poznan.pl/dweiss/rzeczpospolita.

[Woliński2006] Marcin Woliński. 2006. Morfeusz – a Practical Tool for the Morphological Analysis of Polish. In Mieczysław A. Kłopotek, Sławomir T. Wierzchoń, and Krzysztof Trojanowski, editors, *Intelligent Information Processing and Web Mining*, Advances in Soft Computing, pages 503–512. Springer-Verlag.

[Zafar et al.2012] Ayesha Zafar, Afia Mahmood, Farhat Abdullah, Saira Zahid, Sarmad Hussain, and Asad Mustafa. 2012. Developing Urdu WordNet Using the Merge Approach. In *Proc. Conference on Language and Technology*, pages 55–59.

[Zaśko-Zielińska et al.2015] Monika Zaśko-Zielińska, Maciej Piasecki, and Stan Szpakowicz. 2015. A Large Wordnet-based Sentiment Lexicon for Polish. In *Proc. RANLP 2015*, page to appear.

[Zgółkowa1994 2005] Halina Zgółkowa, editor. 1994–2005. *Praktyczny słownik współczesnej polszczyzny [A practical dictionary of contemporary Polish]*. Wydawnictwo Kurpisz.