plWordNet 3.0 – a Comprehensive Lexical-Semantic Resource

Marek Maziarz[A], **Maciej Piasecki**[A], **Ewa Rudnicka**[A],
Stan Szpakowicz[B], **Paweł Kędzia**[A]
[A] Wrocław University of Technology, Wrocław, Poland
[B] University of Ottawa, Ottawa, Ontario, Canada
mawroc@gmail.com, maciej.piasecki@pwr.wroc.pl, ewa.rudnicka78@gmail.com,
szpak@eecs.uottawa.ca, pawel.kedzia@pwr.edu.pl

Abstract

We have released plWordNet 3.0, a very large wordnet for Polish. In addition to what is expected in wordnets – richly interrelated synsets – it contains sentiment and emotion annotations, a large set of multi-word expressions, and a mapping onto WordNet 3.1. Part of the release is enWordNet 1.0, a substantially enlarged copy of WordNet 3.1, with material added to allow for a more complete mapping. The paper discusses the design principles of plWordNet, its content, its statistical portrait, a comparison with similar resources, and a partial list of applications.

1 Introduction

WordNet (Fellbaum, 1998), developed at Princeton University and available on an open licence since the early 1990s, has proven useful in thousands of applications to English texts. It is not flawless, but it strikes a most reasonable balance between the formalisation of the descriptions of lexical meaning and the wide coverage required for practical applications. Wordnets for other languages have been built upon the WordNet blueprint, but almost none of them come close to WordNet's size and coverage. That limits their influence on language technology for those languages. It is therefore unclear whether the success of the "WordNet phenomenon" is not somehow restricted to English. It must also be noted that most of those wordnets have been translated, one way or another, from Princeton WordNet, mainly in order to reduce the workload and cost. This construction method does not quite take into consideration the peculiarities of the given language's lexical semantic systems, inasmuch as the lexical material and the network of relations strongly depend on the solutions specific to English.

The goal of the plWordNet team was to build a wordnet which provides a faithful and comprehensive description of the system of Polish lexical semantics. That is to say, its structure should represent accurately the lexico-semantic relations between lexical meanings in Polish, and be motivated only by observations derived from Polish language data. We were determined to avoid any form of translation from wordnets for other language, and even any kind of structure transfer. That was meant to keep our wordnet's structure free from the idiosyncrasies of the lexical systems of other languages. We also aimed to have a resource with good coverage with respect to lemmas, word senses and instances of lexico-semantic relations, so that the resulting language resource could be a strong basis for practical applications with a high chance of retrieving semantic knowledge. Finally, we assumed that our wordnet should be developed in close correspondence to language data collected from very large corpora, so that it could become a robust, faithful description of Polish usage.

We have been fortunate in the past 10 years to have almost continual funding at a level that allowed us to reach our goals without compromising these fundamental assumptions. It was a rare chance to carry out a long-term plan of building a very large wordnet without worrying too much about cost.[1] The main purpose of this paper is to present plWordNet, to square its final state with the assumptions, and to compare it with several other lexical resources. We will also refer to hundreds of plWordNet's known applications and thus try and show that the effort was worth the price.

2 plWordNet in brief

2.1 The plWordNet model

Wordnets have become standard lexical-semantics resources in NLP, and have found thousands of applications. A wordnet is now considered a basic language resource, expected to be available for any language.

[1] There were three major releases. The development was carried out by researchers (linguists and computational linguists), wordnet editors (supporting linguists) and programmers (developing and maintaining tools to support linguists' work), at the approximate cost of 40 person-years.

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 2259–2268, Osaka, Japan, December 11-17 2016.

The plWordNet project, arising from a wish to fill a gap in language technology for Polish, and clearly inspired by WordNet, aimed to produce a faithful description of the system of Polish lexical semantics.

It must be noted that several fundamental definitions in the WordNet paradigm, *e.g.,* those of a synset, near synonymy or lexicalised concepts, were not clear enough to be used operationally (Fellbaum, 1998; Vossen, 2002), and to achieve good consistency among wordnet editors – see a longer discussion in (Piasecki et al., 2009). We decided against the transfer method (Vossen, 2002), so as to avoid influencing plWordNet's structure with some properties alien to the Polish lexical system. We also could not adopt the merge method, because no dictionaries or other lexical resources on open licenses were available.[2] We proposed a *corpus-based wordnet development process* instead: a large text corpus is a primary data source, and language tools and systems help wordnet editors explore the corpus.[3]

The corpus has been the main knowledge source for all phases of the development, from the systematic extraction of lemmas for inclusion in plWordNet to the automated acquisition of lexico-semantic relations for presentation to the editors. Dictionaries and encyclopaedias complement language competence of the editors, all of them trained linguists, and in all linguistic matters editors have the last word. Detailed instructions ensure a high degree of consistency of those decisions.

Corpora contain words, with senses discernible by context. Groups of synonyms are not a natural phenomenon in texts. We decided to make the *lexical unit* (LU) the basic building block in plWordNet, rather than the synset as in WordNet (Piasecki et al., 2009). We defined the LU in a rather technical way as a triple: a lemma, its part of speech and its sense indicator. We assumed that one LU belongs to exactly one *synset*. The synset, however, has been defined indirectly – and operationally – as a group of LUs which share lexico-semantic *constitutive relations* and *constitutive features* (Maziarz et al., 2013). Examples of the former are hyponymy, hypernymy, meronymy and holonymy; of the latter, stylistic register, aspect, and semantic classes for adjectives and verbs. With this definition of the synset, a relation between two synsets in plWordNet can be treated as a shorthand for the fact that LUs from the two groups share links by certain relation, *e.g.,* hypernymy.

Each relation has been given a clear definition meant to allow wordnet editors to make consistent decisions. There also are linguistic substitution tests, with slots to be occupied by two LUs possibly in this relation. The tests, which support wordnet editors' decisions very effectively, are automatically filled and presented in a wordnet editing system called WordnetLoom (Piasecki et al., 2013). We adhere intentionally to the *minimal commitment principle*: lexico-semantic relations are grounded in the Polish linguistic tradition and language data in very large corpora; plWordNet's structure is derived from the relations in a way which depends on no particular theory of meaning.

2.2 The content

description layer	instances
lexico-semantic relations	>700K
glosses	>100K
usage examples	83K
links to Wikipedia	55K
sentiment annotation	30K

Table 1: Multilayered semantic description in plWordNet: the statistics.

The relations are the backbone of a wordnet: they jointly describe a word's meaning; definitions and usage example come next. plWordNet has over 40 different relation types (100 when counting subtypes). many of them link LUs from different parts of speech. In addition to relations, plWordNet describes meaning in several ways. Table 1 presents the statistics of these descriptions.

- **Semantic domains** (Princeton WordNet calls them *lexicographer files*) are broad lexical fields of a given LU. They are quite general (*e.g.,* `animals`, `artifacts`, `place`).
- **Stylistic labels** describe the lexical register of a given LU. There are 11 registers in plWordNet: non-standard, obsolete, regional, terminological, argot/slang, literary, official, vulgar, coarse, colloquial, general; words in some registers (*e.g.,* vulgar and coarse) can co-exist in a synset, but normally distinct registers mean distinct synsets. The register thus affects the network of lexical relations.

[2]Unrestricted availability was an essential point for us in view of what we wanted plWordNet's licence to be.

[3]The corpus grew in size from the initial ≈260 million words during the work on plWordNet 1.0, through ≈1.8 billion tokens for plWordNet 2.3, to ≈4.0 billion for plWordNet 3.0.

	synsets	lemmas	LUs	avs
GermaNet	101,371	119,231	131,814	–
PWN	117,659	155,593	206,978	1.74
enWN	125,500	165,712	218,611	1.74
plWN	**197,721**	**179,125**	**260,214**	1.32

Table 2: The count of synsets, lemmas and lexical units (LUs), and average synset size (avs), in PWN 3.1 (PWN), enWordNet 1.0 (enWN), plWordNet 3.0 (plWN) and GermaNet 10.0 (`http://www.sfs.uni-tuebingen.de/GermaNet/`).

- **Glosses** are short definitions, a very important element of plWordNet. They help the user to understand the network, and plWordNet editors to work with high effectiveness.
- **Usage examples** are sentences which illustrate a particular lexical meaning. They are exemplars for sense usage and also real corpus evidence. Usage examples in plWordNet are due to the liguists' intuition, or taken from corpora in the public domain or published on a Creative Commons Licence.
- **Links to *Wikipedia*** are added to those LUs whose meaning is an exact equivalent of a Wikipedia entry.
- **Semantic verb classes**, part of plWordNet's structure, generalise the Vendler classes for typical Polish verb usage. They influence the network's shape, since only verbs of the same class may be linked with hyponymy.[4]
- **Sentiment and emotion annotation** marks word meanings as discussed below.

Sentiment analysis or the construction of a sentiment lexicon, perhaps based on plWordNet, has been a frequently stated intended use of plWordNet once it became publicly available.[5] We met this expectation in a pilot project, in which about 30,000 noun and adjective LUs were annotated with basic emotions (Plutchik, 1980), fundamental human values and sentiment polarity, illustrated by usage examples (Zaśko-Zielińska et al., 2015). LUs rather than synsets were annotated, because LUs from the same synset can differ with respect to sentiment polarity.[6] Annotation covers the sentiment polarity of a sense on a 5-level scale, and basic emotions.) and LUs are the object of linguistic tests or are included in usage examples. The annotation was performed by a group separate from the plWordNet editors, so it also served as a form of verification of the plWordNet content.

The newest release of plWordNet, version 3.0, complements the preceding versions. After version 2.3, the work concentrated on a modified system of relations for adjectives (Maziarz et al., 2012) and on the expansion of the adjective sub-database; the construction of the adverb subnetwork,[7] supported by a semi-automated method based on adjective-adverb derivational relations (Maziarz et al., 2016); and a major increase of the number of lexicalised multi-word expressions (Dziob and Wendelberger, 2016).

3 Comparative analysis

3.1 The lexical net

A wordnet is a lexical net, so it can be evaluated with statistical measures suitable for graphs (Lewis, 2009). We consider graph size, network volume, average graph density, corpus coverage, clustering coefficient, distance measure and connectivity. A wordnet of good quality ought to have a large, dense network, covering contemporary corpora well, and showing traits of "small-worldness".

Network volume and density. Table 2 shows the number of synsets, lemmas and LUs in three manually and independently constructed wordnets: Princeton WordNet, plWordNet and GermaNet, together with enWordNet, our extension of Princeton WordNet. We can say that plWordNet is comparable in size to Princeton WordNet (and the 5% larger enWordNet), and almost twice as large as GermaNet. Table 3 shows that the volumes of the two resources are also comparable. plWordNet has 208K LU relation instances and 324K synset relation instances; the WordNet counts are 91K and 195K, respectively. Taking into account that in WordNet the average synset size is higher than the average synset size in plWordNet (Table 2) one may want to calculate an average relation density per LU. This measure approximates an

[4]For example, *zgubić*$_2$ and *stracić*$_1$ 'to lose' (`HAPPPENINGS`) or *wybudować*$_1$ 'build$_{\texttt{PERF}}$' and *zrobić*$_2$ 'do$_{\texttt{PERF}}$' (`PERFECTIVE ACTIONS`).

[5]An independent attempt has been made (Haniewicz et al., 2013; Haniewicz et al., 2014).

[6]For instance, *pies*$_2$ 'Canis lupus familiaris' is unmarked, while *pies*$_3$ 'cop (policeman)' is negatively marked.

[7]Adverbs are usually neglected in wordnet: there are none in GermaNet, and less than 3% of all lexical units in WordNet are adverbs. Their proper relational description is not easy, as witnessed by WordNet's low synset relation density of 0.03 (Table 1).

WordNet 3.1	verbs		nouns		adverbs		adjectives		all	
	N	ρ	N	ρ	N	ρ	N	ρ	N	ρ
LU relations	24,840	0.99	44,185	0.28	720	0.13	21,636	0.72	91,381	0.42
synset relations	16,827	1.22	145,338	1.62	109	0.03	23,491	1.29	185,765	1.48
all relation types	80,280	3.20	492,457	3.12	1,015	0.18	86,221	2.87	659,973	3.02
plWordNet 3.0	**verbs**		**nouns**		**adverbs**		**adjectives**		**all**	
	N	ρ	N	ρ	N	ρ	N	ρ	N	ρ
LU relations	48,744	1.50	98,376	0.58	12,542	1.14	48,894	1.02	208,556	0.80
synset relations	36,616	1.66	219,266	1.75	19,716	2.18	48,258	1.17	323,856	1.64
all relation types	127,065	3.92	494,893	2.94	43,551	3.94	118,574	2.47	784,083	3.02

Table 3: The volume of the lexical networks and relation density with regard to parts of speech. N is the number of relation instances, ρ is the relation density measured either for LUs, or synsets, or for all relation types.

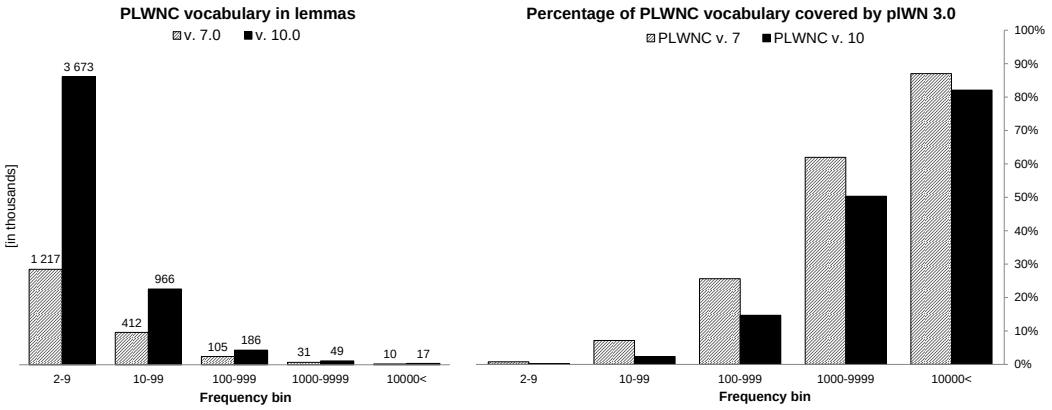

Figure 1: **Left**: The number of lemmas in PLWNC version 3.0 and 10.0 with regard to different freqency bins. The bin "100-999" contains those words that occur in the PLWNC 100 to 999 times. In agreement with Zipf's law, there are far more rare than frequent words in both corpora. **Right**: Coverage of the 7th and 10th version of PLWNC by plWordNet 3.0.

amount of information falling to a single LU, which in fact is very similar for both wordnets, 660K for WordNet and 785K for plWordNet, see row "all relation types" in the table.[8]

Corpus coverage. Figure 1, right, shows how well plWordNet 3.0's vocabulary covers PLWNC. plWordNet was developed on three corpora, the ICS PAS corpus (Przepiórkowski, 2004) (plWordNet 1.0, 250M tokens), plWordNet Corpus 7.0 (plWordNet 2.0 and 3.0, 1.8G tokens) and plWordNet Corpus 10.0 (plWordNet 3.0, 4.2G tokens). Note that the coverage of PLWNC 10.0 is lower than that of version 7.0. The chart also proves that plWordNet creators favoured more frequent lemmas over less frequent. Figure 2, left, presents the coverage of three versions of plWordNet (1.0, 2.0 and 3.0). The consecutive versions of plWordNet housed more and more low-frequenct lemmas. Now, words with frequencies lower than $f = 10$ account for merely 10% of plWordNet 3.0 (Figure 2, right).

Small world. Similarly to Princeton WordNet, plWordNet shows a small-world behaviour: short average path length and high clustering coefficient (Sigman and Cecchi, 2001).[9] In Figure 3 we plot the statistics for three versions of plWordNet (1.0, 2.0, 3.0), Princeton WordNet and a conglomerate, an effect of mapping from plWordNet 3.0 to WordNet 3.1 (WN-plWN3). For a classical random graph of plWordNet's size, a global clustering coefficient is close to $\frac{\langle k \rangle}{N} = 2.5 \times 10^{-5}$, where $\langle k \rangle$ is an average number of neighbours of a vertex (see ρ values in Table 3, we put here $\langle k \rangle = 3$), and N is the number of graph vertices (in this case synsets, see Table 2). The average path length for the random graph is very similar to the obtained values (see see Figure 3): $\frac{ln(N)}{ln(\langle k \rangle)} \approx 11$ (Omidi and Masoudi-Nejad, 2009).

For sure, plWordNet is denser now in terms of the clustering coefficient and the shorter path lengths than in the past (it is indeed a smaller world now). As compared to WordNet, plWordNet versions 2.0 and 3.0 have shorter average path length and higher clustering coefficient.

[8]This approximation was calculated thus: we choose synset relations within the same POS and multiply the number of relation instances by a square of the average synset size for a particular POS (synset relations are shorthand for relations between LUs from two synsets). If a synset relation holds between different POSs, we multiply the number of synset relations by the average synset sizes of the two distinct POSs.

[9]We calculate the classic *global clustering coefficient* (Opsahl, 2013).

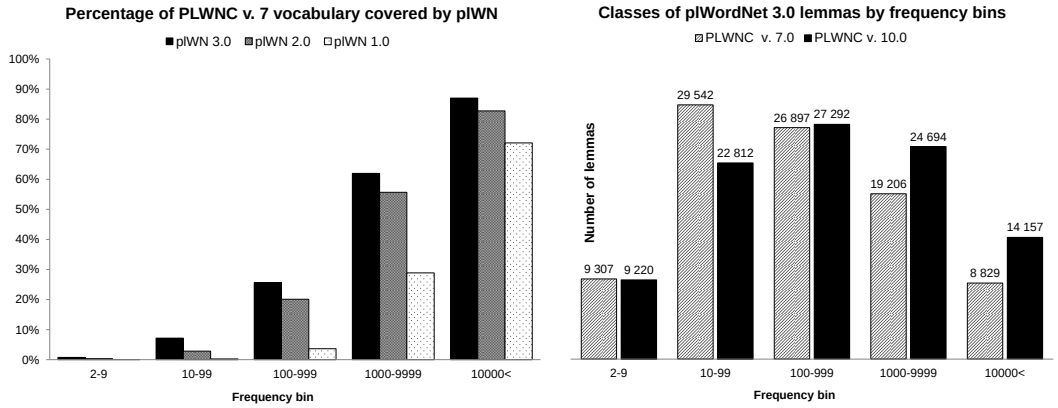

Figure 2: **Left**: Coverage of the 7th version of plWordNet Corpus (PLWNC) by three different stages of plWordNet development – versions 1.0 (from 2009), 2.0 (2013) and 3.0 (2016) – with regard to frequency bins. The bin "100-999" contains words which occur in the PLWNC 100-999 times. Percentages show how many lemmas in each corpus bin are found in plWordNet (version 1st, 2nd or 3rd). **Right**: The cardinality of frequency bins in plWordNet 3.0. Frequencies were calculated in two versions of plWordNet Corpus (7.0 i 10.0).

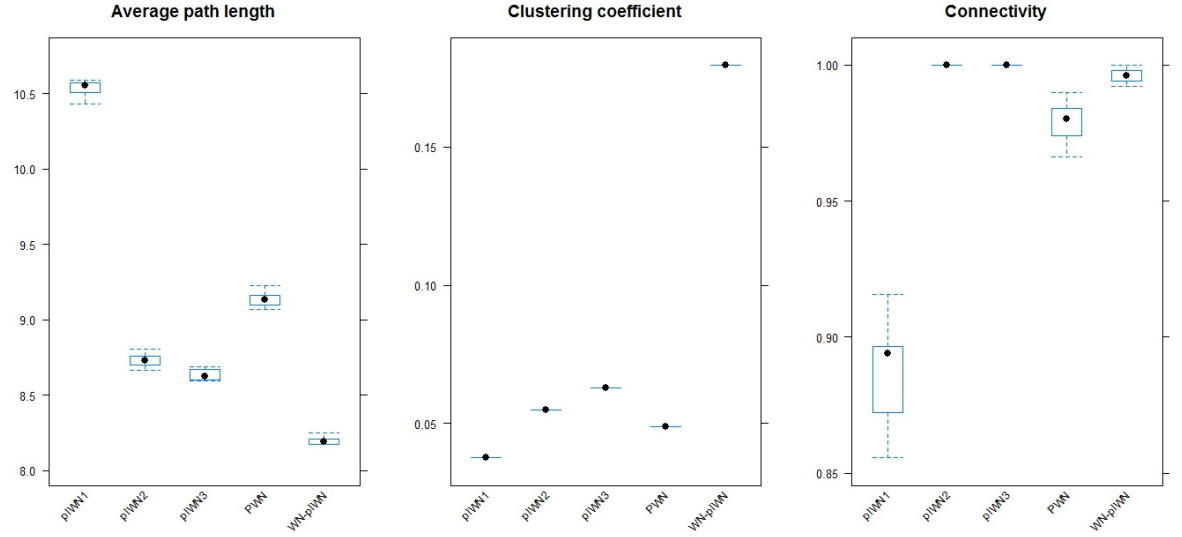

Figure 3: Average path length, clustering coefficient and connectivity in different lexical networks. plWN1, plWN2, plWN3: = plWordNet 1.0, 2.0, 3.0; PWN: WordNet 3.1, WN-plWN: mapping between plWordNet3.0 and WordNet 3.1. Clustering coefficients were calculated for the whole graphs. Average path lengths were obtained by randomly picking a pair of 2×500 synsets (without replacement) and seeking a way through the graph between the pairs; if a way could be found, the shortest path was chosen, and then the set of resulting calculations was averaged. The procedure was repeated 10 times for each graph. The connectivity was calculated simultaneously: it is a ratio of felicitously found paths.

Connectivity measures how often a path can be established between two synsets randomly chosen in a graph. For all wordnet versions, the statistic is high (>85%) or very high (>95%), with plWordNet 1.0 last in ranking and two other versions of plWordNet with the two highest ranks.

The mapping results, described in the next section, were very surprising. The merged networks of Polish and English lexical units gave impressive values of clustering coefficient (3 times larger than for plWordNet 3.0) and shortest path lengths. The conglomerate has small-world behaviour more than its separate parts. It seems that linking independently built resources creates a new quality.

3.2 Comparison by mapping

As noted, plWordNet has been developed independently from WordNet, without any transfer of structures between the two resources, thus avoiding any bias towards WordNet. Even so, the alignment of plWordNet and WordNet was needed for a variety of (bilingual and multilingual) applications and research tasks. We have designed a strategy of mapping plWordNet to WordNet (Rudnicka et al., 2012). The key element

I-relation	Noun	Adjective	Adverb	Total
I-Synonymy	36,367	4,077	448	40,892
I-Hyponymy	74,394	29,216	781	104,391
I-Hypernymy	4,121	167	51	4,339
I-Meronymy	6,982	-	-	6,982
I-Holonymy	3,471	-	-	3,471
I-Partial synonymy	4,339	1,544	4	5,887
I-Inter-register synonymy	1,672	54	22	1,748
I-Cross-categorial synonymy	-	19,286	-	19,286
Total	131,346	54,344	1,306	186,996

Table 4: Interlingual relation counts

of the strategy was a comparison of the two relation structures in order to find the corresponding nodes of synset graph structures and link them via one of eight interlingual relations (hierarchically ordered by varying strength and specificity). The mapping was done manually, in the WordNetLoom editor (Piasecki et al., 2013), bottom-up (leaves first), from plWordNet to WordNet. As a result, almost all plWordNet noun synsets are mapped in version 3.0, about $\frac{3}{4}$ of adjective synsets and about $\frac{1}{4}$ of adverb synsets.

The linguists' work was supported by an automatic prompt system which suggested interlingual links using a rule-based part-of-speech-sensitive algorithm, and a cascade dictionary (Kędzia et al., 2013; Rudnicka et al., 2015a). The final decisions, however, were made by linguists and the cost of the mapping process was comparable to that of editing plWordNet. That has turned out to be money well spent, for two reasons. The two interlinked, independently created wordnets provide a remarkable opportunity to run a comparative analysis; and the mapping process required a careful analysis of plWordNet's structure, so it was a kind of evaluation procedure.

Indeed, the mapping process enabled a comparative analysis and an evaluation of the lexical coverage and the construction methods of the two wordnets. The linguists discovered many gaps in the lexical coverage between plWordNet and Princeton WordNet, as well as numerous differences in the number, type and structure of synset and LU relations – all due to the different construction methods (Rudnicka et al., 2015b). These facts account for the final mapping results, with interlingual hyponymy counts doubling interlingual synonymy counts. This is illustrated in Table 4.

The results are striking. Interlingual synonymy was most highly favoured by the mapping procedure, yet its counts are much lower than those of interlingual hyponymy across all mapped categories. This is caused by the strict restrictions on the application of I-Synonymy. It could only be assigned given strong correspondence of the meanings and relation structures between plWordNet and WordNet synsets. Superficially, noun synset relation structures seem largely to correspond, with hyponymy forming the backbone of a relation network. However, on a closer look, various contrasts come to the fore.

First, plWordNet and WordNet differ in synset granularity, which affects relation structures. In general, plWordNet synsets are smaller and tend to include fewer lexical units than WordNet synsets. In plWordNet there are always distinct synsets for feminine, masculine and neuter forms, singular and plural, mass and count, diminutive, augmentative and stylistically marked forms. While mapping, we found many instances of mixed WordNet synsets grouping together marked and unmarked forms of such pairs. Moreover, the concept of hyponymy in plWordNet and in WordNet is different. plWordNet always understands hyponymy narrowly, as "and hyponymy": the hyponyms have to have all properties of their hypernym(s). That leads to many cases of multiple hyponymy, but it is always of the "and" type. WordNet also allows a more relaxed "or hyponymy", which lets hyponyms have *some* properties of their hypernym(s). We have also found places (both in plWordNet and in WordNet) where the same conceptual dependency was encoded variously by meronymy or by hyponymy.

Adjective and adverb relation structures diverge even more between plWordNet and WordNet than noun relations structures (Rudnicka et al., 2015a). In plWordNet, the adjective synset relation structure is a vertical, hyponymy-based network, partly similar to that for nouns. WordNet employs a completely different, horizontal dumbbell model, based on a rather vague "Similar to" relation. That has made designing an adjective mapping procedure a real challenge. We had to take into account the lexical unit relation network which displays more similarity to establish interlingual correspondence links between plWordNet and WordNet synsets. Since adverbs have been systematically derived from adjectives, we have also capitalised on the results of adjective mapping in designing the mapping procedure for adverbs. The relevant interlingual adjective relation links were copied to adverbs and presented in the form of automatic prompts to linguists. They verified them and introduced manual interlingual adverb links. That process

	plWordNet	WordNet
Nouns	2,733	43,575
Verbs	22,029	13,789
Adjectives	8,188	11,298
Adverbs	7,529	2,704
Total	40,479	71,366

Table 5: The number of synset not mapped yet in plWordNet 3.0 and Princeton WordNet 3.1.

also allowed for critical evaluation (sometimes followed by correction) of interlingual adjective links.

Having finished their work on mapping synsets from selected wordnet graphs (usually domain-restricted), bilingual linguists reported potential errors in plWordNet to the team responsible for the Polish side, who analysed and, if needed, corrected them. Despite meticulous quality control, it is inevitable that isolated errors – typos, flawed links, synsets too general or too specific – persist in plWordNet 3.0. Such errors will be rooted out when a reporting system for users has been implemented.

The mapping went in the usual "national wordnet to WordNet" direction. We were well-aware of substantial lexical, grammatical and cultural differences between English and Polish as well as different development processes of the two wordnets. Even so, we did not expect differences in the mapping coverage between the wordnets as large as those illustrated in Table 5.

The reasons for the discrepancies in the mapping coverage of nouns and adjectives have been already discussed. The mapping of adverbs has only started, while verbs have not been mapped yet.

In short, the results of mapping have shown large differences between plWordNet and WordNet in lexical content, coverage and relation structure. Differences in lexical content are due to lexico-grammatical differences between English and Polish and the existence of many lexical and cultural gaps between the two languages. Differences in lexical coverage are due to different construction methods of the two wordnets: merge method for WordNet and corpus-based method for plWordNet, as well as in the time span of their construction: mid 1990-ties to 2006 for WordNet 3.0 and 2005-2016 for plWordNet 3.0.

The differences in relation structure are due to different theoretical solutions assumed in the construction of two wordnets: lower vs higher synset granularity, "and" vs "or" hyponymy, and the use of hyponymy and meronymy to code the same conceptual distinctions. The effects of those differences are the prevalence of I-hyponymy over I-synonymy and the large part of WordNet not mapped yet, due to one-directional, plWordNet to WordNet mapping direction.

An I-hyponymy-based bilingual resource is clearly less valuable than one based on I-synonymy (due to the lower specificity of links). So, we have sought remedies. One idea was to exploit the existing I-hyponymy links to extend WordNet's coverage. The result was the construction of enWordNet 1.0, an extended version of WordNet. The lemmas of plWordNet leaf synsets linked by I-hyponymy to WordNet synsets were automatically translated by a large cascade dictionary. The obtained list of translations was then filtered by WordNet lemmas. Next, the results of this filtering were divided into lemmas for which the cascade dictionary found: (1) equivalents whose lemmas were not present in WordNet; (2) no equivalents; (3) equivalents whose lemmas were already present in WordNet.

Linguists started with the first group, carefully verifying the suggestions with 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, verifying the existing mapping relations. Moreover, whenever linguists started work with a particular WordNet "nest", they were encouraged to look for its possible extensions on their own (not limiting themselves to cascade dictionary suggestions). The effect of that work is a substantially enlarged version of WordNet, with lexical material – some 10,000 lemmas – added in many places where a link from the Polish side would have been inaccurate. The result, enWordNet 1.0,[10] is also part of this release, which ought to encourage comparative studies and cross-lingual research.

4 Applications of plWordNet

Language resources are developed for applications: the higher the uptake, the better the perceived quality. plWordNet is a pivotal element of a system of language and knowledge resources; plWordNet's wide coverage helps a lot. The system has several layers, with plWordNet in the middle:

- top- and medium-level ontology SUMO with plWordNet semi-automatically mapped onto it (Kędzia and Piasecki, 2014),

[10]The symbol WordNet® is a registered trademark. We cannot use it.

- NELexion2, a very large lexicon of Polish Proper Names (PNs), ≈1.5 million, manually linked at the level of fine-grained semantic PN classes (Marcińczuk, 2016),
- a lexicon of ≈60,000 multiword expressions with syntactic structures described, linked to plWordNet's LUs by lemmas (Maziarz et al., 2015; Dziob et al., 2016),
- a syntactic-semantic lexicon of Polish valency frames (≈15,000 lemmas described) linked to plWordNet at the LU level and semantic restrictions of frame arguments (Kotsyba, 2014; Hajnicz, 2014).

The system is a very large network, linking knowledge elements to lexical meaning and descriptions of local syntactic-semantic structures. Given the mapping to WordNet, the system can be an anchor to a global Linked Data network,[11] a powerful cloud of heterogenous data webs. Manually crafted lexical-semantic resources could serve as a skeleton for the cloud, notably with plWordNet's comprehensive coverage. Lexical item descriptions therein would be the means of anchoring webs to text clouds.

plWordNet has become an important reference for research on the development of wordnets; (Fišer and Sagot, 2015) is the latest of numerous citations.

plWordNet's open license enables frequent use as a monolingual and bilingual dictionary: Web-based (`http://plwordnet.pwr.edu.pl`) via an Android application, and via WordnetLoom (Piasecki et al., 2013) (`http://ws.clarin-pl.eu/public/WordnetLoom-Viewer.zip`) a wordnet editor which offers advanced visual, graph-based browsing. plWordNet has also been included in a very large and popular Polish multilingual dictionary Lingo (`http://ling.pl`). Access to plWordNet as a dictionary amounts to tens of thousand of visits a month.

In addition to monolingual resources, plWordNet is part of multilingual resources, *e.g.,* WordTies (Pedersen et al., 2012), Open Multilingual WordNet (Bond and Foster, 2013) and multimodal resources, *e.g.,* the classification of gestures based on the verb categorisation in plWordNet (Lis and Navarretta, 2014). plWordNet was referred to in the resource for textual entailment (Przepiórkowski, 2015) and utilised for ontology mapping and linking ontology to lexicon (Jastrząb et al., 2016).

Assorted applications of plWordNet include language correction, relation extraction (Mykowiecka and Marciniak, 2014), text indexing (Kaleta, 2014), Text Mining (Maciołek and Dobrowolski, 2013), text classification (Wróbel et al., 2016; Mirończuk and Protasiewicz, 2016), Open Domain Question Answering (Przybyła, 2013), and use as a quasi-ontology in document structure recognition (Kamola et al., 2015).

Registered users of plWordNet declare its applications. Here is a selection of such declaration: education (at different levels) including Polish language teaching, building dictionaries, extraction of synonyms and semantically related words, detection of loanwords, cross-linguistic study on phonesthemes, classification of metaphorical expressions, corpus studies, grammar development, comparative and contrastive studies, language recognition, parsing disambiguation, semantic analysis of text, document similarity measures, semantic indexing of documents, semantic information retrieval, recommendation systems, construction of chatbots and dialogue systems, plagiarism detection, translation evaluation, data visualisation, research on complex networks and ontologies. An exceptional case is the practical use of plWordNet during the medical treatment of aphasia.

5 Always more to do

The release of plWordNet 3.0 is a caesura, but language resources never really reach a stable state. The wordnet is an NLP-friendly description of the Polish lexical system on a scale unheard of even in previously published large unilingual dictionaries.

And yet, each element of the system could stand improvement. For example, while many derivational relations (typical of strongly inflected languages such as Polish) have been introduced, there remains a motherlode of relations signalled by verbal prefixes, a highly productive operation similar to what phrasal verbs contribute to English. Relation density in plWordNet is quite satisfactory, but there can be semi-automatic methods of improving it further. Stylistic registers as a constitutive feature can lead the the natural introduction of sub-databases of specialised vocabulary for a variety of domains, interlinked across registers. Multi-word expressions and proper names need more work. Emotion annotations have to be extended onto the whole network.

Last but not least, user feedback in matters small (typos, omissions) and large (new functionalities, support for new kinds of applications) ought to be implemented.

Acknowledgment: work financed as part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education.

[11] `http://linkeddata.org/`

References

Francis Bond and Ryan Foster. 2013. Linking and Extending an Open Multilingual Wordnet. In *Proc. 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria.

Agnieszka Dziob and Michał Wendelberger. 2016. Extraction and description of multi-word lexical units in plWordNet 3.0. In *Proc. 8th Int. Global Wordnet Conference.*

Agnieszka Dziob, Michał Kaliński, Marek Maziarz, Maciej Piasecki, Adam Radziszewski, Stan Szpakowicz, and Michał Wendelberger. 2016. MWELexicon. Language resource published in CLARIN-PL repository, April.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database.* MIT Press.

Darja Fišer and Benoît Sagot. 2015. Constructing a poor man's wordnet in a resource-rich world. *Language Resources and Evaluation*, 49(3):601–635.

Elżbieta Hajnicz. 2014. Lexico-Semantic Annotation of Składnica treebank by means of plwn lexical units. In *Proc. Seventh Global Wordnet Conference*, pages 23–31, Tartu, Estonia.

Konstanty Haniewicz, Wojciech Rutkowski, Magdalena Adamczyk, and Monika Kaczmarek. 2013. Towards the Lexicon-Based Sentiment Analysis of Polish Texts: Polarity Lexicon. In *Computational Collective Intelligence. Technologies and Applications: 5th International Conference, ICCCI 2013, Craiova, Romania*, pages 286–295. Springer.

Konstanty Haniewicz, Monika Kaczmarek, Magdalena Adamczyk, and Wojciech Rutkowski. 2014. Polarity Lexicon for the Polish Language: Design and Extension with Random Walk Algorithm. In *Advances in Systems Science: Proc. International Conference on Systems Science 2013 (ICSS 2013)*, pages 173–182. Springer.

Tomasz Jastrząb, Grzegorz Kwiatkowski, and Paweł Sadowski. 2016. Mapping of Selected Synsets to Semantic Features. In *Beyond Databases, Architectures and Structures. Advanced Technologies for Data Mining and Knowledge Discovery: 12th International Conference, BDAS 2016, Ustroń, Poland*, pages 357–367. Springer.

Zbigniew Kaleta. 2014. Semantic text indexing. *Computer Science*, Vol. 15 (1):19–34.

Grzegorz Kamola, Michal Spytkowski, Mariusz Paradowski, and Urszula Markowska-Kaczmar. 2015. Image-based logical document structure recognition. *Pattern Analysis and Applications*, 18(3):651–665.

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.

Natalia Kotsyba. 2014. Using Polish Wordnet for Predicting Semantic Roles for the Valency Dictionary of Polish Verbs. In *Advances in Natural Language Processing: 9th International Conference on NLP, PolTAL 2014, Warsaw, Poland*, pages 202–207. Springer.

Pawel Kędzia and Maciej Piasecki. 2014. Ruled-based, Interlingual Motivated Mapping of plWordNet onto SUMO ontology. In *Proc. Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland*, pages 4351–4358.

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

Magdalena Lis and Costanza Navarretta. 2014. Classifying the Form of Iconic Hand Gestures from the Linguistic Categorization of Co-occurring Verbs. In *Proc. 1st European Symposium on Multimodal Communication University of Malta; Valletta; October 17-18; 2013*, volume 101 of *Linköping Electronic Conference Proceedings*, pages 41–50. Linköping University Electronic Press.

Przemysław Maciołek and Grzegorz Dobrowolski. 2013. Cluo: web-scale text mining system for open source intelligence purposes. *Computer Science*, Vol. 14 (1)(1):45–62.

Michał Marcińczuk. 2016. NELexicon2. Language resource – lexicon of Polish Proper Names – published in CLARIN-PL repository, April.

Marek Maziarz, Stanisław Szpakowicz, and Maciej Piasecki. 2012. Semantic Relations among Adjectives in Polish WordNet 2.0: A New Relation Set, Discussion and Evaluation. *Cognitive Studies*, 12:149–179.

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

Marek Maziarz, Stan Szpakowicz, and Maciej Piasecki. 2015. A Procedural Definition of Multi-word Lexical Units. In *Proc. RANLP'2015*, pages 427–435, Hissar, Bulgaria.

Marek Maziarz, Stan Szpakowicz, and Michał Kaliński. 2016. Adverbs in plWordNet: Theory and Implementation . In *Proc. of GWC 2016*, pages 209–217.

Marcin Mirończuk and Jarosław Protasiewicz. 2016. A Diversified Classification Committee for Recognition of Innovative Internet Domains. In *Beyond Databases, Architectures and Structures. Advanced Technologies for Data Mining and Knowledge Discovery: 12th International Conference, BDAS 2016, Ustroń, Poland*, pages 368–383. Springer.

Agnieszka Mykowiecka and Małgorzata Marciniak. 2014. Attribute Value Acquisition through Clustering of Adjectives. In *Advances in Natural Language Processing: 9th International Conference on NLP, PolTAL 2014, Warsaw, Poland*, pages 92–104, Cham. Springer.

Saeed Omidi and Ali Masoudi-Nejad, 2009. *Computational Social Network Analysis: Trends, Tools and Research Advances*, chapter Network Evolution: Theory and Mechanisms, pages 191–224. Springer Science & Business Media.

Tore Opsahl. 2013. global clustering coefficient. *Social Networks*, 35(2).

Bolette Sandford Pedersen, Lars Borin, Markus Forsberg, Krister Lindén, Heili Orav, and Eirikur Rögnvalssson. 2012. Linking and Validating Nordic and Baltic Wordnets- A Multilingual Action in META-NORD. In *Proc. 6th International Global Wordnet Conference*.

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.

Maciej Piasecki, Michał Marcińczuk, Radosław Ramocki, and Marek Maziarz. 2013. WordNetLoom: a WordNet development system integrating form-based and graph-based perspectives. *International Journal of Data Mining, Modelling and Management*, 5(3):210–232.

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

Adam Przepiórkowski. 2015. Towards a Linguistically-Oriented Textual Entailment Test-Suite for Polish Based on the Semantic Syntax Approach. *Cognitive Studies / Études Cognitive*, 15:177–191.

Adam Przepiórkowski. 2004. *The IPI PAN Corpus, Preliminary Version*. Institute of Computer Science PAS.

Piotr Przybyła. 2013. Question Classification for Polish Question Answering. In *Proc. of IIS 2013*, pages 50–56, Warsaw, Poland.

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.

Ewa Rudnicka, Wojciech Witkowski, and Michał Kaliński. 2015a. A Semi-automatic Adjective Mapping Between plWordNet and Princeton WordNet. In *Text, Speech, and Dialogue*, pages 360–368. Springer.

Ewa Rudnicka, Wojciech Witkowski, and Michał Kaliński. 2015b. Towards the Methodology for Extending Princeton WordNet. *Cognitive Studies*, 15(15):335–351.

Mariano Sigman and Guillermo A. Cecchi. 2001. Global organization of the Wordnet lexicon. In *Proc. National Academy of Sciences of the United States of America*, volume 99.

Piek Vossen. 2002. EuroWordNet. Technical report, Univ. of Amsterdam.

Krzysztof Wróbel, Maciej Wielgosz, Aleksander Smywiński-Pohl, and Marcin Pietron. 2016. Comparison of SVM and Ontology-Based Text Classification Methods. In *Proc. of ICAISC 2016*, pages 667–680, Zakopane, Poland. Springer.

Monika Zaśko-Zielińska, Maciej Piasecki, and Stan Szpakowicz. 2015. A Large Wordnet-based Sentiment Lexicon for Polish. In *Proc. RANLP 2015*, pages 721–730.